Concurrencia y Distribución

2015/2016

Tercero, Grado

Dr. Arno Formella

Departamento de Informática
Escola Superior de Enxeñaría Informática
Universidade de Vigo

15/16

profesorado (teoría, últimas prácticas)

Profesor: Arno FORMELLA
Web: http://formella.webs.uvigo.es
Correo: formella@uvigo.es
Tutorías Lu: 09:30–13:30 y 16:30–18:30

usamos: FAITIC/TEMA

profesorado (prácticas)

Profesor: Francisco Javier RODRÍGUEZ MARTÍNEZ
Correo: franjrm@uvigo.es
Tutorías Lu: 13:00–15:00 y 16:00–17:30

Profesor: David OLIVIERI CECCHI
Correo: olivieri@uvigo.es
Tutorías Lu: 10:00–14:00 y 16:00–18:00

Profesor: Mª Lourdes BORRAJO DIZ
Correo: lborrajo@uvigo.es
Tutorías Ma: 10:00–13:00 y Mi: 10:00–13:00

tutorías

- Cambios puntuales de tutorías via aviso web (yo en mi página principal)
- Idiomas: Galego, Castellano, English, Deutsch
- Las transparencias serán en castellano.

horas de dedicación (planificación según guía)

	pres.	no-pres.	suma
Actividades introductorias	0.50	–	0.50
Sesión magistral	18.00	9.00	27.00
Estudios/actividades previos	–	17.00	17.00
Prácticas en aulas de informática	26.00	26.00	52.00
Resolución de problemas y/o exercicios	1.50	19.50	21.00
Presentacións/exposicións	–	1.75	1.75
Tutoría en grupo	1.25	1.25	2.50
Pruebas de respuesta corta	1.00	–	1.00
Pruebas de respuesta larga	2.00	–	2.00
Informes/memorias de prácticas	–	12.00	12.00
Probas prácticas	1.00	–	1.00
Resolución de problemas y/o exercicios	–	12.00	12.00
Otras	0.25	–	0.25
Suma	51.50	98.50	150.00

horas presenciales

Teoría:	los miércoles, 11:00-12:30 horas, Aula 2.2
Prácticas:	6 grupos, Lab. 37 a partir del martes 02/02/2015
	2 sem. Fran, 5 sem. David, 6 sem. Arno

3º CURSO - 2C	LUNES	MARTES	MIÉRCOLES	JUEVES	VIERNES
09:00 - 09:30 09:30 - 10:00 10:00 - 10:30 10:30 - 11:00	SI_4 [SO2] HAE_1 [TecElect]	CDI_2 [L37] SI_3 [SO2] DGP_1 [L38]	TALF_2 [L31B] DGP_4 [L38] CDI_5 [L37]	DGP_5 [L38] HAE_3 [TecElect] SI_1 [SO2]	CDI_3 [L37] HAE_2 [TecElect] TALF_5 [L31B]
11:00 - 11:30 11:30 - 12:00 12:00 - 12:30	HAE Carlos Castro Miguéns [2.2]	DGP Celso Campos Bastos [2.2]	CDI Arno Formella [2.2]	SI Juan Carlos González Moreno [2.2]	TALF Manuel Vilares Ferro [2.2]
12:30 - 13:00 13:00 - 13:30 13:30 - 14:00 14:00 - 14:30	SI_2 [SO2] HAE_5 [TecElect]	CDI_4 [L37] SI_5 [SO2]	TALF_1 [L31B] DGP_3 [L38]	DGP_2 [L38] TALF_4 [L31B]	CDI_1 [L37] HAE_4 [TecElect]
15:00 – 15:30 15:30 – 16:00					
16:00 – 16:30 16:30 – 17:00 17:00 - 17:30 17:30 - 18:00	SI_6 [SO2]			TALF_3 [L31B]	TALF_6 [L31B]
18:00 - 18:30 18:30 – 19:00 19:00 – 19:30 19:30 – 20:00		CDI_6 [L37]			

horas presenciales

Luns	Martes	Mércores	Xoves	Venres	Sábado	Domingo
			21/ene	22/ene	23/ene	24/ene
25/ene	26/ene	27/ene	28/ene	29/ene	30/ene	31/ene
01/feb	02/feb	03/feb	04/feb	05/feb	06/feb	07/feb
08/feb	09/feb	10/feb	11/feb	12/feb	13/feb	14/feb
15/feb	16/feb	17/feb	18/feb	19/feb	20/feb	21/feb
22/feb	23/feb	24/feb	25/feb	26/feb	27/feb	28/feb
29/feb	01/mar	02/mar	03/mar	04/mar	05/mar	06/mar
07/mar	08/mar	09/mar	10/mar	11/mar	12/mar	13/mar
14/mar	15/mar	16/mar	17/mar	18/mar	19/mar	20/mar
21/mar	22/mar	23/mar	24/mar	25/mar	26/mar	27/mar
28/mar	29/mar	30/mar	31/mar	01/abr	02/abr	03/abr
04/abr	05/abr	06/abr	07/abr	08/abr	09/abr	10/abr
11/abr	12/abr	13/abr	14/abr	15/abr	16/abr	17/abr
18/abr	19/abr	20/abr	21/abr	22/abr	23/abr	24/abr
25/abr	26/abr	27/abr	28/abr	29/abr	30/abr	01/may
02/may	03/may	04/may	05/may	06/may	07/may	08/may
09/may	10/may	11/may	12/may	13/may	14/may	15/may
16/may	17/may	18/may	19/may	20/may	21/may	22/may
23/may	24/may	25/may	26/may	27/may	28/may	29/may
30/may	31/may					

horas en aula presenciales

- 27.01. actividad introductoria
- 13 sesiones magistrales + pruebas de respuesta corta
- 19.05. prueba final (10:00-14:00, 2 horas)
- 04.07. prueba terminal (15:30-19:00, 2 horas)
- 13 sesiones prácticas (menos los grupos de los viernes)

horas de trabajo (este curso)

Actividad	pres.		no-pres.	horas	guía
Actividades introductorias	0.5	T	–	0.5	(0.5)
Sesión magistral	19.5	T	7.5	27.0	(27.0)
Estudios/actividades previos	–		6.5	6.5	(17.0)
Prácticas en aulas de informática	25.0	P	26.0	51.0	(52.0)
Resolución de problemas/exercicios	1.5	T	19.5	21.0	(21.0)
Presentacións/exposicións	–	P	1.0	1.0	(1.7)
Tutoría en grupo	1.3		1.0	2.3	(2.5)
Pruebas de respuesta corta	1.0	T	–	1.0	(1.0)
Pruebas de respuesta larga	2.0		–	2.0	(2.0)
Informes/memorias de prácticas	–		12.0	12.0	(12.0)
Probas prácticas	1.0	P		1.0	(1.0)
Resolución de problemas y/o exercicios	–		12.0	12.0	(12.0)
Otras	0.2		–	0.2	(0.3)
Suma	52.0		85.5	137.5	
segunda conv.			+12.5	+12.5	150.0

+ –

horas del profesor (yo, aproximado, optimista)

960	horas anuales
	$\cdot 101.5/240$
406	-21 horas presenciales teoría
385	-21 horas preparación clases teoría
364	$-(6\cdot 6\cdot 2-4)=68$ horas presenciales prácticas
296	-12 horas preparación/gestión clases prácticas
284	$-96/4-5$ horas corrección exámenes teoría
265	$-96/2/4*3$ horas corrección entregas prácticas
229	$/96/16$ media por estudiante por semana
0.149	horas
9	minutos por estudiante por semana

prerrequisitos

- matemáticas
- algoritmos y estructuras de datos
- todo sobre programación
- arquitectura de computadoras
- redes
- sistemas operativos
- lenguajes de programación (Java, C++)

contexto en el plan de estudio

ORGANIZACIÓN TEMPORAL DEL TÍTULO DE GRADUADO/A EN INGENIERÍA INFORMÁTICA

SEMESTRES							
1/1S	1/2S	2/1S	2/2S	3/1S	3/2S	4/1S	4/2S
DERECHO:: FUNDAMENTOS ÉTICOS Y JURÍDICOS DE LAS TIC (CFB; 6 ECTS)	EMPRESA:: ADMINISTRACIÓN DE LA TECNOLOGÍA Y LA EMPRESA (CFB; 6 ECTS)	INGENIERÍA DE SOFTWARE I (OB; 6 ECTS)	INGENIERÍA DE SOFTWARE II (OB; 6 ECTS)	INTERFACES DE USUARIO (OB; 6 ECTS)	DIRECCIÓN Y GESTIÓN DE PROYECTOS (OB; 6 ECTS)	OPTATIVA (6 ECTS)	OPTATIVA (6 ECTS)
MATEMÁTICAS:: FUNDAMENTOS MATEMÁTICOS PARA LA INFORMÁTICA (CFB; 6 ECTS)	MATEMÁTICAS:: ÁLGEBRA LINEAL (CFB; 6 ECTS)	MATEMÁTICAS:: ESTADÍSTICA (CFB; 6 ECTS)	BASES DE DATOS I (OB; 6 ECTS)	BASES DE DATOS II (OB; 6 ECTS)	SISTEMAS INTELIGENTES (OB; 6 ECTS)	OPTATIVA (6 ECTS)	OPTATIVA (6 ECTS)
MATEMÁTICAS:: ANÁLISIS MATEMÁTICO (CFB; 6 ECTS)	INFORMÁTICA:: ALGORITMOS Y ESTRUCTURAS DE DATOS I (CFB; 6 ECTS)	ALGORITMOS ESTRUCTURAS DE DATOS II (OB; 6 ECTS)	REDES DE COMPUTADORAS I (OB; 6 ECTS)	REDES DE COMPUTADORAS II (OB; 6 ECTS)	CONCURRENCIA Y DISTRIBUCIÓN (OB; 6 ECTS)	SEGURIDAD EN SISTEMAS INFORMÁTICOS (OB; 6 ECTS)	TÉCNICAS DE COMUNICACIÓN Y LIDERAZGO (OB; 6 ECTS)
INFORMÁTICA:: PROGRAMACIÓN I (CFB; 6 ECTS)	PROGRAMACIÓN II (OB; 6 ECTS)	SISTEMAS OPERATIVOS I (OB; 6 ECTS)	SISTEMAS OPERATIVOS II (OB; 6 ECTS)	LÓGICA PARA LA COMPUTACIÓN (OB; 6 ECTS)	TEORÍA DE AUTÓMATAS Y LENGUAJES FORMALES (OB; 6 ECTS)	APRENDIZAJE BASADO EN PROYECTOS (OB; 6 ECTS)	TRABAJO FIN DE GRADO (OB; 12 ECTS)
FÍSICA:: SISTEMAS DIGITALES (CFB; 6 ECTS)	INFORMÁTICA:: ARQUITECTURA DE COMPUTADORAS I (CFB; 6 ECTS)	ARQUITECTURA DE COMPUTADORAS II (OB; 6 ECTS)	ARQUITECTURAS PARALELAS (OB; 6 ECTS)	CENTROS DE DATOS (OB; 6 ECTS)	HARDWARE DE APLICACIÓN ESPECÍFICA (OB; 6 ECTS)	OPTATIVA (6 ECTS)	
30 ECTS	30 ECTS	30 ECTS	30 ECTS	30 ECTS	30 ECTS	30 ECTS	30 ECTS

evaluación continua

- (P1) preguntas cortas
  [A4 A5 A7 A8 A12 A13 A14 A15 A16 A19 A20 A22 A25 A26 A27 A28 A30 A31 A33 A35 A36 B1 B2 B5 B6 B8 B10]
- (P2) preguntas largas (hay que aprobar ≥ 4)
  [A4 A5 A7 A8 A12 A13 A14 A15 A16 A19 A20 A22 A25 A26 A27 A28 A30 A31 A33 A35 A36 B1 B2 B5 B6 B8 B10]
- (P3) informes
  [A4 A5 A7 A8 A12 A13 A14 A15 A16 A19 A20 A22 A25 A26 A27 A28 A30 A31 A33 A35 A36 B1 B2 B3 B5 B6 B8 B10]
- (P4) programación (hay que aprobar ≥ 4)
  [A4 A5 A7 A8 A12 A13 A14 A15 A16 A19 A20 A22 A25 A26 A27 A28 A30 A31 A33 A35 A36 B1 B2 B5 B6 B8 B10]
- (P5) análisis
  [A4 A5 A7 A8 A12 A13 A14 A15 A16 A19 A20 A22 A25 A26 A27 A28 A30 A31 A33 A35 A36 B1 B2 B5 B6 B8 B10]
- (P6) presentaciones
  [A4 A5 A7 A8 A12 A13 A14 A15 A16 A19 A20 A22 A25 A26 A27 A28 A30 A31 A33 A35 A36 B1 B2 B3 B5 B6 B8 B10]

- $\min(10, \min(5, 0.2P_1 + 0.4P_2) + \min(4, 0.25P_3 + 0.25P_4) + 0.05P_5 + 0.05P_6) \geq 5$

evaluación no-asistentes

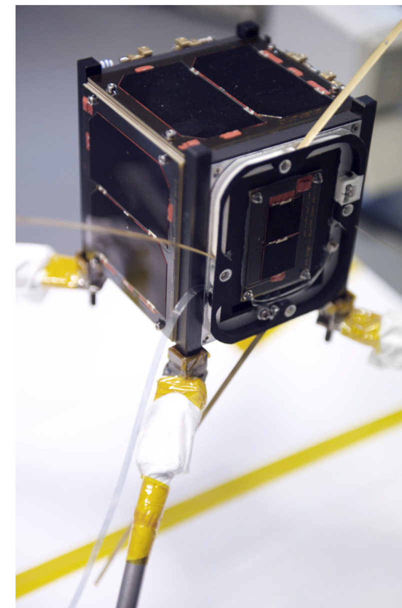
- examen (jueves, 19.05.) de 3.5 horas (entre 10:00-13:30) que cubre todo el contenido de la asignatura (teoría y prácticas)
- y/o examen (lunes, 04.07.) de 3.5 horas (entre 15:30-19:00) que cubre todo el contenido de la asignatura (teoría y prácticas)
- alumnos del curso puente tendrán ciertas consideraciones especiales (quedan por determinar según demanda)
- un alumno o bien se autodeclara no-asistente o lo muestra por no asistir a por lo menos 80 % de las actividades presenciales (como mucho se puede faltar a 9 horas de las (21+24=45) horas de presencialidad principal)

vos y nos

¿Quién soy yo?

- http://formella.webs.uvigo.es
- http://lia.ei.uvigo.es

al espacio...



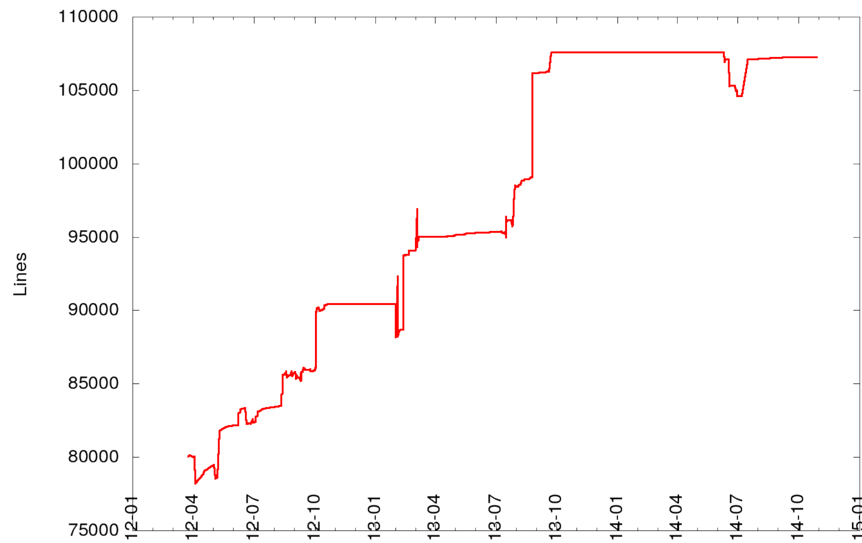
3 Satélites:

XaTcobeo, 14/02/2012

HumSAT, 21/11/2013

Serpens, 20/08–17/09/2015

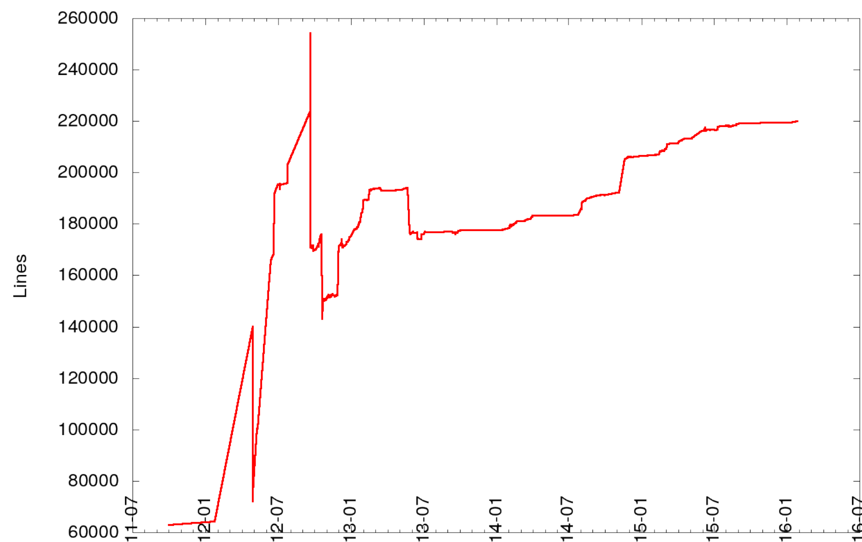
HumSAT software evolution (lines of code)



reconocimiento de formas

- investigación y desarrollo en el ámbito de reconocimiento de formas
  - reconocer formas
  - aprender formas
- uso en aplicaciones de
  - educación infantil,
  - educación matemática,
  - interfaces amigables
  - interfaces para motores de búsqueda
- miramos un poco el prototipo Shapewiz...

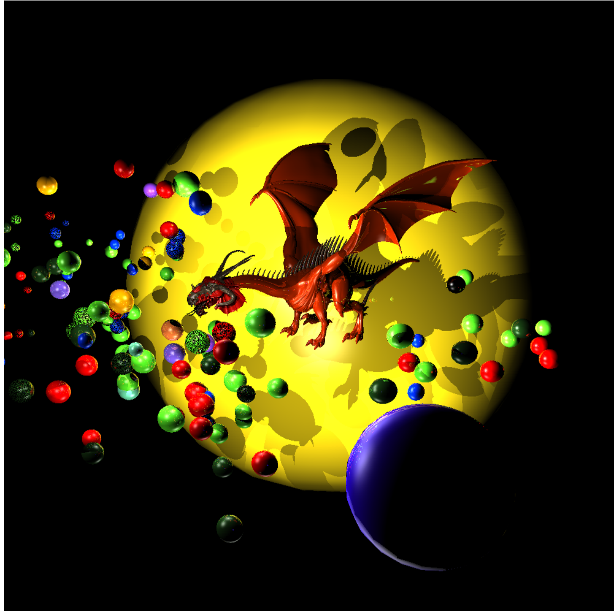
Shapewiz software evolution (lines of code)



trazado de rayos

- descripción de un modelado de una escena (objecto, colores, texturas, luces, etc.)
- simulación de una cámara digital
- obtención de una imágen *foto-realista*

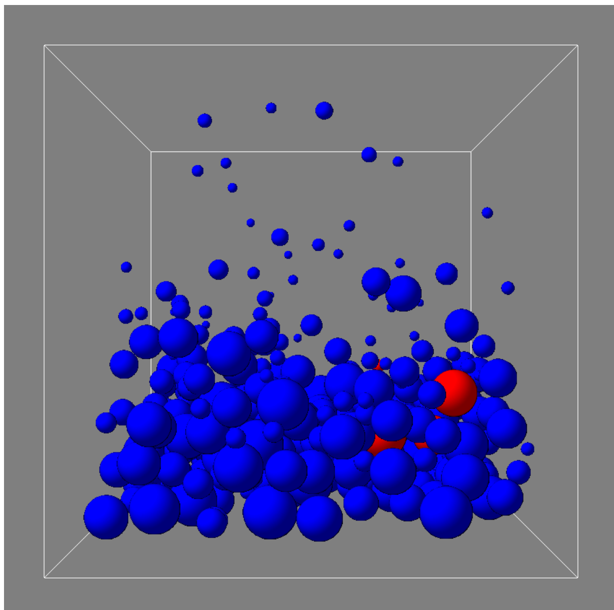
trazado de rayos



simulación de partículas

- descripción de un modelado de partículas
  (tamaño, propiedades pre/pos-colisión, otros objetos, etc.)
- simulación según modelado física
  (p.ej. con gravitación)
- simulación basado en eventos discretos
- estudio de efectos en materiales granulares y gases

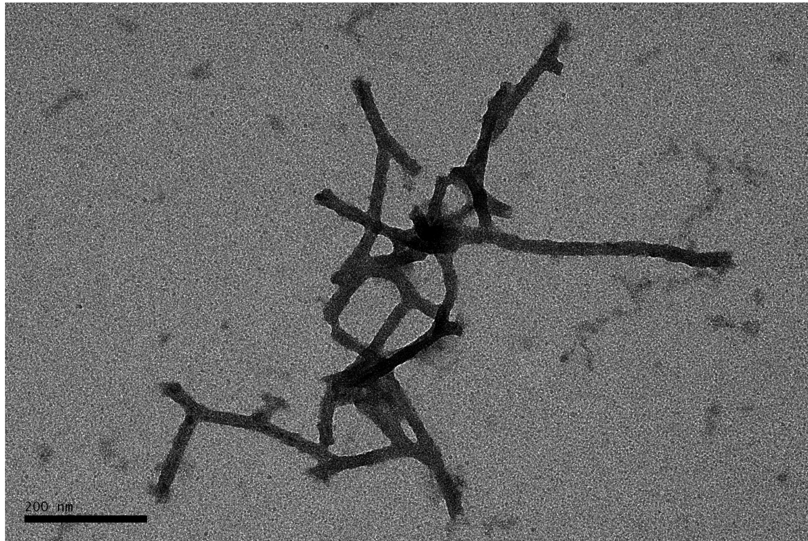
pemd



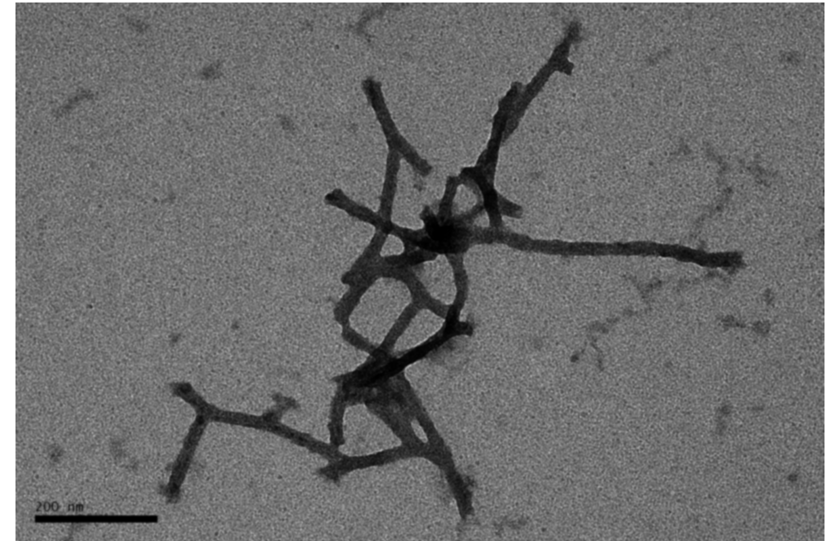
pemd

```
pemd_g -Ri 0.5 -Ra 3.0 -c 1000000 -en 0.75 -i
cdi.pemd -g 2 -pn 343 -ps 343
```

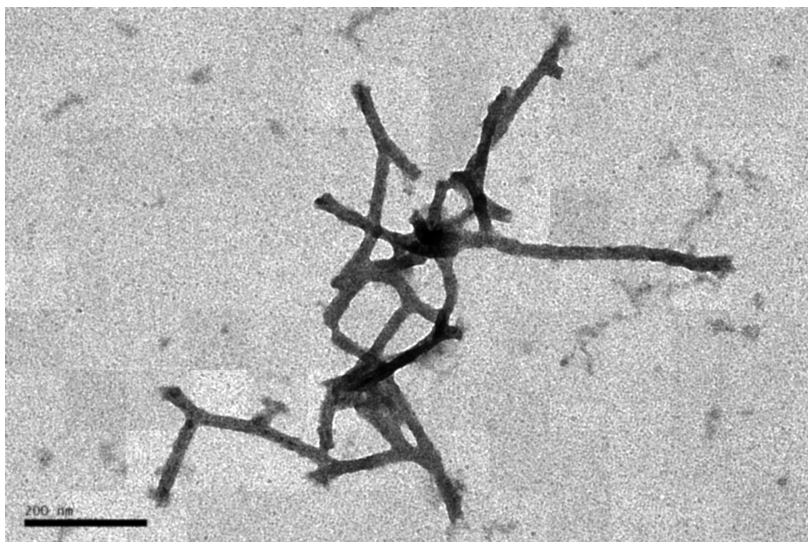
bio-nanotubos (original)



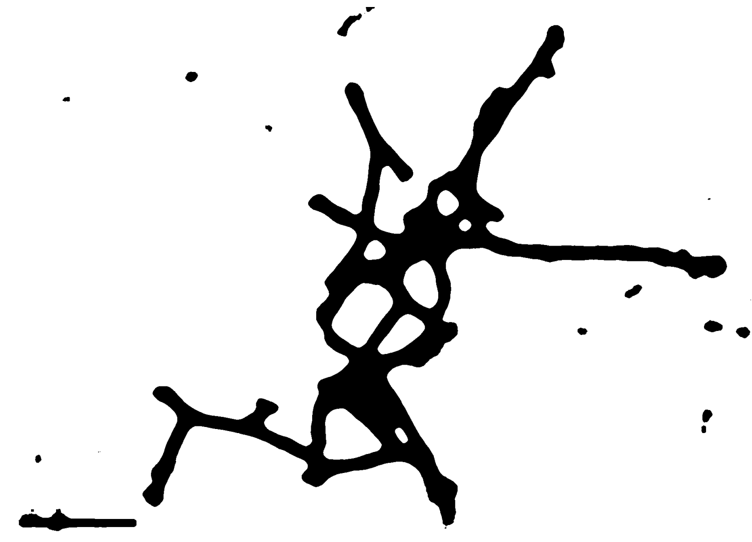
bio-nanotubos (menos ruido)



bio-nanotubos (más contraste)

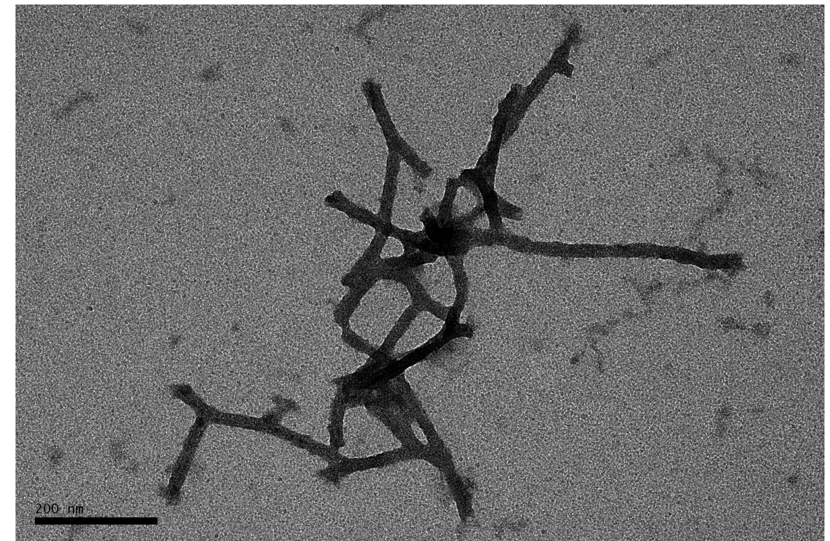


bio-nanotubos (binarizado)



procesamiento de imágenes
bio-nanotubos (esqueleto)

procesamiento de imágenes
bio-nanotubos (original)



posible colaboración, ejemplos durante el curso

- Todas estas aplicaciones se usarán como vehículos de ejemplo para mostrar aspectos de concurrencia y distribución.
- Quien quiere colaborar, por ejemplo en su TFG, puede ponerse en contacto conmigo.

contenido (optimista)

Tema	**Contenido**
Sistemas concurrentes y distribuidos	Concepto de la programación concurrente y distribuida, Introducción al modelado de sistemas concurrentes y distribuidos, Arquitecturas hardware para la concurrencia y distribución, Herramientas para del desarrollo de aplicaciones concurrentes y distribuidos
Procesos	Concepto de procesos, Planificador, Atomicitad y exclusión mutua, Concurrencia transactional, Reloj y estado distribuido

contenido (optimista)

Tema	Contenido
Sincronización y comunicación	Sincronización y comunicación en sistemas concurrentes y distribuidos, Sincronización y comunicación a nivel bajo y alto, Seguridad y vivacidad en sistemas concurrentes y distribuidos
Herramientas de programación y desarrollo de aplicaciones	Programación concurrente y distribuida con JAVA (y C/C++), Patrones de diseño para el desarrollo de aplicaciones concurrentes y distribuidos, Herramientas y metodologías de diseño, verificación y depuración de aplicaciones concurrentes y distribuidos

Nota

Prácticamente todas las asignaturas optativas en uno u otro aspecto requieren del concepto de concurrencia y distribución en sistemas modernos para lograr sus objetivos específicos.

documento de transparencias

- Este documento crecerá durante el curso,
  ojo, no necesariamente solamente al final.
- Habrá más documentos (capítulos de libros, manuales, etc.) con que trabajar durante el curso.
- Los ejemplos de programas y algoritmos serán en inglés.
- Las transparencias no están (posiblemente/probablemente) ni correctos ni completos.

competencias, un intento...

Competencias	Tipo	Cod.
Que os estudantes demostren posuír e comprender coñecementos nunha área de estudo que parte da base da educación secundaria xeral e adoita atoparse a un nivel que, malia se apoiar en libros de texto avanzados, inclúe tamén algúns aspectos que implican coñecementos procedentes da vangarda do seu campo de estudo.	facer	CB1
Que os estudantes saiban aplicar os seus coñecementos ó seu traballo ou vocación dunha forma profesional e posúan as competencias que adoitan demostrarse por medio da elaboración e defensa de argumentos e a resolución de problemas dentro da súa área de estudo.	facer	CB2
Que os estudantes teñan a capacidade de reunir e interpretar datos relevantes (normalmente dentro da súa área de estudo) para emitir xuízos que inclúan unha reflexión sobre temas relevantes de índole social, científica ou ética.	facer	CB3
Que os estudantes desenvolvan aquelas habilidades de aprendizaxe necesarias para emprender estudos posteriores cun alto grao de autonomía.	facer	CB4

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade para concebir, redactar, organizar, planificar, desenvolver e asinar proxectos no ámbito da enxeñaría en informática que teñan por obxecto, de acordo cos coñecementos adquiridos , a concepción, o desenvolvemento ou a explotación de sistemas, servizos e aplicacións informáticas.	facer	CG1
Capacidade para dirixir as actividades obxecto dos proxectos do ámbito da informática de acordo cos coñecementos adquiridos.	saber	CG2
Capacidade para deseñar, desenvolver, avaliar e asegurar a accesibilidade, ergonomía, usabilidade e seguridade dos sistemas, servizos e aplicacións informáticas, así como da información que xestionan.	facer	CG3
Capacidade para definir, avaliar e seleccionar plataformas hardware e software para o desenvolvemento e a execución de sistemas, servizos e aplicacións informáticas, de acordo cos coñecementos adquiridos.	facer	CG4

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade para concebir, desenvolver e manter sistemas, servizos e aplicacións informáticas empregando os métodos da enxeñería de software como instrumento para o aseguramento de sua calidade, de acordo cos coñecementos adquiridos.	facer	CG5
Capacidad para concebir e desenvolver sistemas ou arquitecturas informáticas centralizadas ou distribuidas integrando hardware, software e redes de acordo cos coñecementos adquiridos.	facer	CG6
Capacidade para coñecer, comprender e aplicar a lexislación necesaria durante o desenvolvemento da profesión de Enxeñeiro Técnico en Informática e manexar especificacións, regulamentos e normas de obrigado cumprimento.	saber	CG7
Coñecemento das materias básicas e tecnoloxías, que capaciten para a aprendizaxe e desenvolvemento de novos métodos e tecnoloxías, así como as que lles doten dunha gran versatilidade para adaptarse a novas situacións.	facer	CG8

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade para resolver problemas con iniciativa, toma de decisións, autonomía e creatividade. Capacidade para saber comunicar e transmitir os coñecementos, habilidades e destrezas da profesión de Enxeñeiro Técnico en Informática.	saber	CG9
Capacidade para analizar e valorar o impacto social e medioambiental das solucións técnicas, comprendendo a responsabilidade ética e profesional da actividade de Enxeñeiro Técnico en Informática.	saber	CG11
Coñecemento e aplicación de elementos básicos de economía e de xestión de recursos humáns, organización e planificación de proxectos, así como a lexislación, regulación e normalización no ámbito dos proxectos informáticos, de acordo cos coñecementos adquiridos.	saber	CG12

competencias, un intento...

Competencias	Tipo	Cod.
Coñecementos básicos sobre o uso e programación dos ordenadores, sistemas operativos, bases de datos e programas informáticos con aplicación na enxeñería	facer	CE4
Coñecemento da estrutura, organización, funcionamento e interconexión dos sistemas informáticos, os fundamentos da súa programación, e a súa aplicación para a resolución de problemas propios da enxeñería	saber	CE5
Capacidade para deseñar, desenvolver, seleccionar e avaliar aplicacións e sistemas informáticos, asegurando a súa fiabilidade, seguridade e calidade, conforme aos principios éticos e á lexislación e normativa vixente	saber	CE7

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade para planificar, concibir, despregar e dirixir proxectos, servizos e sistemas informáticos en tódolos ámbitos, liderando a súa posta en marcha e mellora continua e valorando o seu impacto económico e social	saber	CE8
Coñecemento e aplicación dos procedementos algorítmicos básicos das tecnoloxías informáticas para deseñar solucións a problemas, analizando a idoneidade e complexidade dos algoritmos propostos	facer	CE12

competencias, un intento...

Competencias	Tipo	Cod.
Coñecemento, deseño e utilización de forma eficiente dos tipos e estruturas de datos máis axeitados á resolución dun problema	facer	CE13
Capacidade para analizar, deseñar, construír e manter aplicacións de forma robusta, segura e eficiente, elixindo o paradigma e as linguaxes de programación máis axeitadas	facer	CE14
Capacidade de coñecer, comprender e avaliar a estrutura e arquitectura dos computadores, así como os compoñentes básicos que os conforman	facer	CE15

competencias, un intento...

Competencias	Tipo	Cod.
Coñecemento das características, funcionalidades e estrutura dos Sistemas Operativos e deseñar e implementar aplicacións baseadas nos seus servizos	saber	CE16
Coñecemento e aplicación das ferramentas necesarias para o almacenamento, procesamento e acceso aos Sistemas de información, incluídos os baseados en web	saber	CE19
Coñecemento e aplicación dos principios fundamentais e técnicas básicas da programación paralela, concurrente, distribuída e de tempo real	facer	CE20

competencias, un intento...

Competencias	Tipo	Cod.
Coñecemento e aplicación dos principios, metodoloxías e ciclos de vida da enxeñería de software	saber	CE22
Capacidade para desenvolver, manter e avaliar servizos e sistemas software que satisfagan todos os requisitos do usuario e se comporten de forma fiable e eficiente, sexan asequibles de desenvolver e manter e cumpran normas de calidade, aplicando as teorías, principios, métodos e prácticas da Enxeñería do Software	saber	CE25

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade para valorar as necesidades do cliente e especificar os requisitos software para satisfacer estas necesidades, reconciliando obxectivos en conflito mediante a procura de compromisos aceptables dentro das limitacións derivadas do custo, do tempo, da existencia de sistemas xa desenvolvidos e das propias organizacións	saber	CE26
Capacidade de dar solución a problemas de integración en función das estratexias, estándares e tecnoloxías dispoñibles	saber	CE27
Capacidade de identificar e analizar problemas e deseñar, desenvolver, implementar, verificar e documentar solucións software sobre a base dun coñecemento axeitado das teorías, modelos e técnicas actuais	facer	CE28

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade para deseñar solucións apropiadas nun ou máis dominios de aplicación utilizando métodos da enxeñería do software que integren aspectos éticos, sociais, legais e económicos	saber	CE30
Capacidade para comprender a contorna dunha organización e as súas necesidades no ámbito das tecnoloxías da información e as comunicacións	saber	CE31
Capacidade para empregar metodoloxías centradas no usuario e a organización para o desenvolvemento, avaliación e xestión de aplicacións e sistemas baseados en tecnoloxías da información que aseguren a accesibilidade, ergonomía e usabilidade dos sistemas	saber	CE33

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade para seleccionar, despregar, integrar e xestionar sistemas de información que satisfagan as necesidades da organización, cos criterios de custo e calidade identificados	saber	CE35
Capacidade de concibir sistemas, aplicacións e servizos baseados en tecnoloxías de rede, incluíndo Internet, web, comercio electrónico, multimedia, servizos interactivos e computación móbil	saber	CE36

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade de análise, síntese e avaliación	ser	CT1
Capacidade de organización e planificación	ser	CT2
Comunicación oral e escrita na lingua nativa	ser	CT3
Capacidade de abstracción: capacidade de crear e utilizar modelos que reflictan situacións reais	ser	CT5
Capacidade de deseñar e realizar experimentos sinxelos e analizar e interpretar os seus resultados	ser	CT6
Capacidade de buscar, relacionar e estruturar información proveniente de diversas fontes e de integrar ideas e coñecementos	ser	CT7
Resolución de problemas	ser	CT8

competencias, un intento...

Competencias	Tipo	Cod.
Capacidade de tomar decisións	ser	CT9
Capacidade para argumentar e xustificar loxicamente as decisións tomadas e as opinións	ser	CT10
Capacidade de actuar autonomamente	ser	CT11
Capacidade de traballar en situacións de falta de información e/ou baixo presión	ser	CT12
Capacidade de relación interpersoal	ser	CT15
Razoamento crítico	ser	CT16
Aprendizaxe autónoma	ser	CT18
Creatividade	ser	CT20
Ter iniciativa e ser resolutivo	ser	CT22
Ter motivación pola calidade e a mellora continua	ser	CT24

Bibliografía (de interés) I

1. J.T. Palma Méndez, M.C. Garrido Carrera, F. Sánchez Figueroa, A. Quesada Arencibia. *Programación Concurrente.* Thomson, ISBN 84-9732-184-7, 2003.
2. G. Coulouris, J. Dollimore, T. Kindberg. *Sistemas Distribuidos, Conceptos y Diseño.* Addison Wesley, ISBN 84-7829-049-4, 2001.
3. M.L. Liu. *Computación Distribuida* Peason/Addison Wesley, ISBN 84-7829-066-4, 2004.
4. C. Breshears. *The Art of Concurrency.* O'Reilly, ISBN 978-0-596-52153-0, 2009.

Bibliografía (de interés) II

5. M. Herlihy, N. Shavit. *The Art of multiprocessor programming.* Elsevier-Morgan Kaufmann Publishers, ISBN 978-0-12-370591-4, 2008 (978-0-12-397337-5, 2012 revised edition).
6. D. Schmidt, M. Stal, H. Rohnert, F. Buschmann. *Pattern-Oriented Software Architecture, Pattern for Concurrent and Networked Objects.* John Wiley & Sons, ISBN 0-471-60695-2, 2000.

Bibliografía (Java)

1. K. Arnold et.al. *The Java Programming Language.* Addison-Wesley, 3rd Edition, ISBN 0-201-70433-1, 2000.
2. B. Eckel. *Piensa en Java.* Prentice Hall, 2002.
3. D. Lea. *Programación Concurrente en Java.* Addison-Wesley, ISBN 84-7829-038-9, 2001.

- cualquier libro sobre Java que cubre programación con hilos

Bibliografía (antigua)

1. M. Ben-Ari. *Principles of Concurrent and Distributed Programming.* Prentice-Hall, ISBN 0-13-711821-X, 1990.
2. G.R. Andrews. *Concurrent Programming: Principles and Practice.* Benjamin/Cummings, 1991.
3. J.C. Baeten and W.P. Wiejland. *Process Algebra.* Cambridge University Press, 1990.
4. A. Burns and G. Davies. *Concurrent Programming.* Addison-Wesley, 1993.
5. C. Fencott. *Formal Methods for Concurrency.* Thomson Computer Press, 1996.
6. M. Henning, S. Vinoski. *Programación Avanzada en CORBA con C++.* Addison Wesley, ISBN 84-7829-048-6, 2001.

Bibliografía (más antigua) I

6. C.A.R. Hoare. *Communicating Sequential Processes.* Prentice-Hall, 1985.
7. R. Milner. *Concurrency and Communication.* Prentice-Hall, 1989.
8. R. Milner. *Semantics of Concurrent Processes.* in J. van Leeuwen (ed.), Handbook of Theoretical Computer Science. Elsevier and MIT Press, 1990.
9. J.E. Pérez Martínez. *Programación Concurrente.* Editorial Rueda, ISBN 84-7207-059-X, 1990.
10. A.W. Roscoe. *The Theory and Practice of Concurrency.* Prentice-Hall, 1997.

Bibliografía (on–line)

- Apuntes de esta asignatura:
  http://formella.webs.uvigo.es/doc/cdg15/index.html
- Concurrency JSR-166 Interest Site (antiguado)
  http://gee.cs.oswego.edu/dl/concurrency-interest/index.html
- El antiguo paquete de Doug Lea que funciona con Java 1.4 (antiguado)
  http://formella.webs.uvigo.es/doc/concurrent.tar.gz
  (.tar.gz [502749 Byte])
- The Java memory model (antiguado)
  http://www.cs.umd.edu/%7Epugh/java/memoryModel

Bibliografía VI (adicional)

- G. Bracha. *Generics in the Java Programming Language.* July 5, 2004.
- buscadores en la red

introducción

Existen diversas definiciones de los términos en la literatura:

- programación concurrente
- programación paralela
- programación distribuida

definición

Una posible distinción (según mi opinión) es:

- la programación concurrente se dedica más a desarrollar y aplicar conceptos para el uso de recursos en paralelo (desde el punto de vista de varios actores)
- la programación en paralelo se dedica más a solucionar y analizar problemas bajo el concepto del uso de recursos en paralelo (desde el punto de vista de un sólo actor)

definición

Otra posibilidad de separar los términos es:

- un programa concurrente define las acciones que se pueden ejecutar simultaneamente, es decir, están en progreso
- un programa paralelo es un programa concurrente diseñado de ser ejecutado en hardware paralelo
- un programa distribuido es un programa paralelo diseñado de ser ejecutado en hardware distribuido, es decir, donde varios procesadores no tengan memoria compartida, tienen que intercambiar la información mediante de transmisión de mensajes/datos.

intuición

Intuitivamente, todos tenemos una idea básica de lo que significa el concepto de concurrencia.

ordenamos

©http://kartenspiel.org/knack-kartenspiel/

ordenamos

- Imaginamos que somos dos ordenadores.
- Cada uno tiene muchos procesadores que se pueden comunicar entre ellos.
- Cada procesos tiene un neipe.
- Queremos que al final la barraja esté ordenado.

- ¿Cómo procedemos?
- ¿Cuáles son los aspectos/problemas a tratar?
- ¿Cómo lo trasladamos a un entorno programable?

ordenamos

Creative Common: Sönke Kraft aka Arnulf zu Linden

ordenamos

Creative Common: Sönke Kraft aka Arnulf zu Linden

sumamos

3482 0984	8473 8093	3746 6112	4958 6432
9923 7463	4398 7329	8746 0302	9823 4326
9821 3234	8464 5643	3745 2854	7734 6511
6534 7732	2907 0238	2985 5328	7334 6532
3982 6452	4328 9231	8439 4431	8374 4721
3274 8549	3278 8192	7843 1723	7364 1323
8329 0123	1212 8322	4133 7742	1232 9234
6434 6012	3823 7213	7438 7439	3284 2328

problemas

¿Con qué problemas nos enfrentamos?

- selección del algoritmo
- RAE: Conjunto ordenado (?!) y finito (?!) de operaciones que permite hallar la solución de un problema.
- división del trabajo
- distribución de los datos
- sincronización necesaria
- comunicación de los resultados

problemas

¡Y también con!

- medición de características
- depuración del programa
- (fiabilidad de los componentes)
- (fiabilidad de la comunicación)
- (detección de la terminación)

sumamos otra vez

3482 0984	8473 8093	3746 6112	4958 6432
9923 7463	4398 7329	8746 0302	9823 4326
9821 3234	8464 5643	3745 2854	7734 6511
6534 7732	2907 0238	2985 5328	7334 6532
3982 6452	4328 9231	8439 4431	8374 4721
3274 8549	3278 8192	7843 1723	7364 1323
8329 0123	1212 8322	4133 7742	1232 9234
6434 6012	3823 7213	7438 7439	3284 2328

resultado

3482 0984	8473 8093	3746 6112	4958 6432
9923 7463	4398 7329	8746 0302	9823 4326
9821 3234	8464 5643	3745 2854	7734 6511
6534 7732	2907 0238	2985 5328	7334 6532
3982 6452	4328 9231	8439 4431	8374 4721
3274 8549	3278 8192	7843 1723	7364 1323
8329 0123	1212 8322	4133 7742	1232 9234
6434 6012	3823 7213	7438 7439	3284 2328
5 1783 0549	3 6888 4261	4 7078 5931	5 0107 1407
			18 5857 2148

Java

Este repaso a Java no es

- ni completo
- ni exhaustivo
- ni suficiente

para programar en Java.
Debe servir solamente para refrescar conocimiento ya adquerido y para animar de profundizar el estudio del lenguaje con otras fuentes, por ejemplo, con la bibliografía añadida y los manuales correspondientes (mirad boletín de prácticas).

Java

- Se destacan ciertas diferencias con C++ (otro lenguaje de programación orientado a objetos importante).
- Se comentan ciertos detalles del lenguaje que muchas veces no se perciben a primera vista.
- Se introducen los conceptos ya instrínsicos de Java para la programación concurrente.

hola mundo

El famoso *hola mundo* se programa en Java así:

```
class Hello {
  public static void main(String[] args) {
    System.out.println("Hello world");
  }
}
```

El programa principal se llama `main()` y tiene que ser declarado público y estático. No devuelve ningún valor (por eso se declara como `void`).
Los parámetros de la línea de comando se pasan como un vector de cadenas de letras (`String`).

¿Qué se comenta?

Existen varias posibilidades de escribir comentarios:

`//`	comentario de línea
`/// ...`	comentario de documentación
`/* ... */`	comentario de bloque
`/** ... */`	comentario de documentación

- Se usa doxygen (o javadoc, u otro bueno) para generar automáticamente la documentación. Todos tienen unos comandos para aumentar la documentación.
- Se documenta sobre todo lo que no es obvio, las interfaces (en el sentido amplio de la palabra), y los casos límite.
- Es decir: Los comentarios son las respuestas a preguntas del *¿Cómo?* y del *¿Por qué?*.

inicialización

- Los objetos en Java siempre tienen valores conocidos, los objetos, es decir, sus miembros siempre están inicializados.
- Si el programa no da una inicialización explícita, Java asigna el valor cero, es decir, `0`, `0.0`, `\u0000`, `false` o `null` dependiendo del tipo de la variable.
- Variables locales hay que inicializar antes de usarlas, el código se ejecuta cuando la ejecución llega a este punto.

Java, C++, C#

- Java y C++ (o C#) son, hasta cierto punto, bastante parecidos. (por ejemplo, en su síntaxis y gran parte de sus metodologías), aunque también existen grandes diferencias (por ejemplo, en su no–uso o uso de punteros y la gestión de memoria).
- Se resaltará algunos de las diferencias principales entre Java y C++.

modificadores de clases I

Se pueden declarar clases con uno o varios de los siguientes modificadores para especificar ciertas propiedades (no existen en C++):

- `public` la clase es visible desde fuera del fichero
- `abstract` la clase todavía no está completa, es decir, no se puede instanciar objetos antes de que se hayan implementado en una clase derivada los métodos que faltan
- `final` no se puede extender la clase
- `strictfp` obliga a la máquina virtual a cumplir el estándar de IEEE para los números flotantes

tipos simples II

- float y double son solamente casi-iguales en Java y C++.
- No existen enteros sin signos en Java (pero si en C++).
- Los tipos simples no son clases, pero existen para todos los tipos simples clases que implementan el comportamiento de ellos.
- Desde Java 5 la conversión de tipos simples a sus objetos correspondientes (y vice versa) es automático.
- Sólo hace falta *escribirles* con mayúscula (con la excepción de Integer).
- Las clases para los tipos simples proporcionan también varias constantes para trabajar con los números (por ejemplo, NEGATIVE_INFINITY etc.).

modificadores de acceso

- private: accesible solamente desde la propia clase
- package: (o ningún modificador) accesible solamente desde la propia clase o dentro del mismo paquete
- protected: accesible solamente desde la propia clase, dentro del mismo paquete, o desde clases derivadas
- public: accesible siempre cuando la clase es visible
- (En C++, por defecto, los miembros son privados, mientras en Java los miembros son, por defecto, del paquete.)

modificadores de miembros I

Modificadores de miembros siendo instancias de objetos:

- final: declara constantes si está delante de tipos simples (diferencia a C++ donde se declara constantes con const), aunque las constantes no se pueden modificar en el transcurso del programa, pueden ser calculadas durante sus construcciónes; las variables finales, aún declaradas sin inicialización, tienen que obtener sus valores como muy tarde en la fase de construcción de un objeto de la clase
- static: declara miembros de la clase que pertenecen a la clase y no a instancias de objetos, es decir, todos los objetos de la clase acceden a lo mismo

modificadores de miembros II

- transient: excluye un miembro del proceso de conversión en un flujo de bytes si el objeto se salva al disco o se transmite por una conexión (no hay en C++)
- volatile: ordena a la máquina virtual de Java que no use ningún tipo de cache para el miembro, así es más probable (aunque no garantizado) que varios hilos vean el mismo valor de una variable; declarando variables del tipo long o double como volatile aseguramos que las operaciones básicas sean atómicas (este tema veremos más adelante más en detalle)

modificadores de métodos I

Modificadores de miembros siendo métodos:

- `abstract`: el método todavía no está completo, es decir, no se puede instanciar objetos antes de que se haya implementado el método en una clase derivada (parecido a los métodos puros de C++)
- `static`: el método pertenece a la clase y no a un objeto de la clase, un método estático puede acceder solamente miembros estáticos
- `final`: no se puede sobreescribir el método en una clase derivada (no hay en C++)
- `synchronized`: el método pertenece a una región crítica del objeto (no hay en C++)

modificadores de métodos II

- `native`: propone una interfaz para llamar a métodos escritos en otros lenguajes, su uso depende de la implementación de la máquina virtual de Java (no hay en C++, ahí se realiza durante el linkage)
- `strictfp`: obliga a la máquina virtual a cumplir el estándar de IEEE para los números flotantes (no hay en C++, ahí depende de las opciones del compilador)

marcas I

Adicionalmente Java proporciona `break` con una marca que se puede usar para salir en un salto de varios bucles anidados.

```
mark:
  while(...) {
    for(...) {
      break mark;
    }
  }
```

Es decir, se puede escribir código no-estructurado.

marcas II

- También existe un `continue` con marca que permite saltar al principio de un bucle más alla del actual.
- No existe el `goto` (pero es una palabra reservada), su uso habitual en C++ se puede emular con los `break`s y `continue`s y con las secuencias `try-catch-finally`.

operadores I

Java usa los mismos operadores que C++ con las siguientes excepciones:

- existe adicionalmente `>>>` como desplazamiento a la derecha llenando con ceros a la izquierda
- existe el `instanceof` para comparar tipos (C++ tiene un concepto parecido con `typeid`)
- los operadores de C++ relacionados a punteros no existen
- no existe el `delete` de C++
- no existe el `sizeof` de C++

operadores II

- La prioridad y la asociatividad son las mismas que en C++.
- Hay pequeñas diferencias entre Java y C++, si ciertos símbolos están tratados como operadores o no (por ejemplo, los `[]`).
- Además Java no proporciona la posibilidad de sobrecargar operadores
  (con la excepción de cadenas (`String`) donde el `+` y el `+=` ya están sobrecargados por defecto).

palabras reservadas I

Las siguientes palabras están reservadas en Java:

```
abstract default if         private      this
boolean  do      implements protected    throw
break    double  import     public       throws
byte     else    instanceof return       transient
case     extends int        short        try
catch    final   interface  static       void
char     finally long       strictfp     volatile
class    float   native     super        while
const    for     new        switch
continue goto    package    synchronized
```


palabras reservadas II

- Además las palabras `null`, `false` y `true` que sirven como constantes no se pueden usar como nombres.
- Aunque `goto` y `const` aparecen en la lista arriba, no se usan en el lenguaje.

construcción por defecto

- Sólo si una clase no contiene ningún constructor Java propone un constructor por defecto que tiene el mismo modificador de acceso que la clase.
- Constructores pueden lanzar excepciones como cualquier otro método.

constructores

- Para facilitar la construcción de objetos, es posible usar bloques de código sin que pertenezcan a constructores.
- Esos bloques están prepuestos (en su orden de aparencia) delante de los códigos de todos los constructores.
- El mismo mecanismo se puede usar para inicializar miembros estáticos poniendo un `static` delante del bloque de código.
- Inicializaciones estáticas no pueden lanzar excepciones chequeadas (ya que no se pueden envolver en bloques `try-catch`).

inicialización estática

```
class ... {
  ...
  static int[] powertwo=new int[10];
  static {
    powertwo[0]=1;
    for(int i=1; i<powertwo.length; ++i)
      powertwo[i]=powertwo[i-1]<<1;
  }
  ...
}
```


recolector de memoria

- No existe un operador para eliminar objetos del montón, eso es tarea del recolector de memoria incorporado en Java (diferencia con C++ donde se tiene que liberar memoria con `delete` explícitamente).
- Para dar pistas de ayuda al recolector se puede asignar `null` a una referencia indicando al recolector que no se va a referenciar con esta referencia dicho objeto nunca jamás.
- Antes de ser destruido se ejecuta el método `finalize()` del objeto (por defecto no hace nada).
- En un programa concurrente obviamente el recolector de memoria tiene que ser un algoritmo sobre una estructura de datos concurrente.

parámetros no modificables no existen

- No se puede evitar posibles modificaciones de un parámetro (que sí se puede evitar hasta cierto punto en C++ declarando el parámetro como `const`-referencia).
- Declarando el parámetro como `final` solamente protege la propia referencia (paso por valor).
- La declaración se puede/suele usar como indicación al usuario que se pretende no cambiar el objeto (aunque el compilador no lo garantiza).

vectores

- Los vectores se declaran solamente con su límite súperior dado que el límite ínferior siempre es cero (`0`).
- El código

```
int[] vector = new int[15]
```

  crea un vector de números enteros de longitud 15.

control de acceso

- Java comprueba si los accesos a vectores con índices quedan dentro de los límites permitidos (diferencia con C++ donde no hay una comprobación).
- Si se detecta un acceso fuera de los límites se produce una excepción `IndexOutOfBoundsException`.
- Dependiendo de las capacidades del compilador eso puede resultar en una pérdida de rendimiento.

vectores son objetos

- Los vectores son objetos implícitos que siempre conocen sus propias longitudes (`values.length`) (diferencia con C++ donde tal vector no es nada más que un puntero) y que se comportan como clases finales.
- No se pueden declarar los elementos de un vector como constantes (como es posible en C++), es decir, el contenido de los componentes siempre se puede modificar en un programa en Java.
- No se puede modificar el tamaño de un vector una vez creado (difrencia con C++ donde tal longitud es dinámico).

`this` and `super`

- Cada objeto tiene por defecto una referencia llamada `this` que proporciona acceso al propio objeto (diferencia a C++ donde `this` es un puntero).
- Obviamente, la referencia `this` no existe en métodos estáticos.
- Cada objeto (menos la clase `object`) tiene una referencia a su clase superior llamada `super` (diferencia a C++ donde no existe, se tiene acceso a las clases superiores por otros medios).
- `this` y `super` se pueden usar especialmente para acceder a variables y métodos que están escondidos por nombres locales.

más sobre constructores

- Para facilitar las definiciones de constructores, un constructor puede llamar en su primer sentencia
  - o bien a otro constructor con `this(...)`
  - o bien a un constructor de su superclase con `super(...)` (ambos no exiten en C++).
- El constructor de la superclase sin parámetros se llama en todos los casos al final de la posible cadena de llamadas a constructores `this()` en caso que no haya una llamada explícita.

orden de construcción

La construcción de objetos sigue siempre el siguiente orden:

- construcción de la superclase, nota que no se llama ningún constructor por defecto que no sea el constructor sin parámetros
- ejecución de todos los bloques de inicialización
- ejecución del código del constructor

extender clases

- No se puede extender al mismo tiempo de más de una clase superior (diferencia a C++ donde se puede derivar de más de una clase).
- Se pueden sobreescribir métodos de la superclase.
- Si se ha sobreescrito una cierta función, las demás funciones con el mismo nombre (pero diferente signatura, es decir, tipos de paramétros) siguen visibles desde la clase derivada (en C++ eso no es el caso).
- Dicho último aspecto puede provocar sorpresas... ¿Cuáles?

la clase `Object` I

Todos los objetos de Java son extensiones de la clase `Object`. Los métodos públicos y protegidos de esta clase son

- `public boolean equals(Object obj)` compara si dos objetos son iguales, por defecto un objeto es igual solamente a si mismo
- `public int hashCode()` devuelve (con alta probabilidad) un valor distinto para cada objeto
- `protected Object clone() throws CloneNotSuportedException` devuelve una copia binaria del objeto (incluyendo sus referencias)

la clase `Object` II

- `public final Class getClass()` devuelve el objeto del tipo `Class` que representa dicha clase durante la ejecución
- `protected void finalize() throws Throwable` se usa para finalizar el objeto, es decir, el administrador de la memoria ejecuta este código especial antes de que se libere la memoria
- `public String toString()` devuelvo una cadena describiendo el objeto
- Las clases derivadas deben sobreecribir los métodos adecuadamente, por ejemplo el método `equals`, si se requiere una comparación binaria.

interfaces

- Usando `interface` en vez de `class` se define una interfaz a una clase sin especificar el código de los métodos.
- Una interfaz no es nada más que una especificación de cómo algo debe ser implementado para que se pueda usar en otro código.
- Una interfaz solo puede tener declaraciones de objetos que son constantes (`final`) y estáticos (`static`).
- En otras palabras, todas las declaraciones de objetos dentro de interfaces automáticamente son finales y estáticos, aunque no se haya descrito explícitamente.

interfaces y herencia

- Igual que las clases, las interfaces pueden incorporar otras clases o interfaces.
- También se pueden extender interfaces.
- Nota que es posible extender una interfaz a partir de más de una interfaz:

```
interface ThisOne extends ThatOne, OtherOne { ... }
```

métodos de interfaces

- Todos los métodos de una interfaz son implícitamente públicos y abstractos, aunque no se haya descrito ni `public` ni `abstract` explícitamente (y eso es la convención).
- Los demás modificadores no están permitidos para métodos en interfaces.
- Para generar un programa todas las interfaces usadas tienen que tener sus clases que las implementen.

resumen: interfaces

Las interfaces se comportan como clases totalmente abstractas, es decir,

- no tienen miembros no–estáticos,
- nada diferente a público,
- y ningún código no–estático.

try'n'catch

- Para facilitar la programación de casos excepcionales Java usa el concepto de lanzar excepciones.
- Una excepción es una clase predefinida y se accede con la sentencia

```
try { ... }
catch (SomeExceptionObject e) { ... }
catch (AnotherExceptionObject e) { ... }
finally { ... }
```


orden de ejecución I

- El bloque `try` contiene el código normal por ejecutar.
- Un bloque `catch(ExceptionObject)` contiene el código excepcional por ejecutar en caso de que durante la ejecución del código normal (que contiene el bloque `try`) se produzca la excepción del tipo adecuado.
- Pueden existir más de un (o ningún) bloque `catch` para reaccionar directamente a más de un (ningún) tipo de excepción.
- Hay que tener cuidado en ordenar las excepciones correctamente, es decir, las más específicas antes de las más generales.

orden de ejecución II

- El bloque `finally` se ejecuta siempre una vez terminado o bien el bloque `try` o bien un bloque `catch` o bien una excepción no tratada o bien antes de seguir un `break`, un `continue` o un `return` hacia fuera de la sentencia `try-catch-finally`.

construcción de clases de excepción

Normalmente se extiende la clase `Exception` para implementar clases propias de excepciones, aún también se puede derivar directamente de la clase `Throwable` que es la superclase (interfaz) de `Exception` o de la clase `RuntimeException`.

```
class MyException extends Exception {
  public MyException() { super(); }
  public MyException(String s) { super(s); }
}
```


declaración de excepciones lanzables

- Entonces, una excepción no es nada más que un objeto que se crea en el caso de aparición del caso excepcional.
- La clase principal de una excepción es la interfaz `Throwable` que incluye un `String` para mostrar una línea de error legible.
- Para que un método pueda lanzar excepciones con las sentencias `try-catch-finally`, es imprecindible declarar las excepciones posibles antes del bloque de código del método con `throws ....`

```
public void myfunc(...) throws MyException {...}
```

- En C++ es al revés, se declara lo que se puede lanzar como mucho.

propagación de excepciones

- Durante la ejecución de un programa se propagan las excepciones desde su punto de aparición subiendo las invocaciones de los métodos hasta que se haya encontrado un bloque `catch` que se ocupa de tratar la excepción.
- En el caso de que no haya ningún bloque responsable, la excepción será tratada por la máquina virtual con el posible resultado de abortar el programa.

lanzar excepciones

- Se pueden lanzar excepciones directamente con la palabra `throw` y la creación de un nuevo objeto de excepción, por ejemplo:

```
throw new MyException("this is an exception");
```

- También los constructores pueden lanzar excepciones que tienen que ser tratados en los métodos que usan dichos objetos construidos.

excepciones de ejecución

- Además de las excepciones así declaradas existen siempre excepciones que pueden ocurrir en qualquier momento de la ejecución del programa, por ejemplo, `RuntimeException` o `Error` o `IndexOutOfBoundException`.
- La ocurrencia de dichas excepciones refleja normalmente un flujo de control erróneo del programa que se debe corrigir antes de distribuir el programa a posibles usuarios.

excepciones de ejecución

- Recomendación: Se usan excepciones solamente para casos excepcionales, es decir, si pasa algo no esperado.
- Excepciones en programas concurrentes se convierten rápidamente en pesadillas.
- Mira Safe (Disciplined) Exception Handling principle con sus dos posibilidades de acción
  - fallo: re-establecer una invariante necesaria
  - re-intentar: re-hacer la operación con (quizá) cierta modificación

Hilos
objetivos

Se usan los hilos para ejecutar varias secuencias de instrucciones de modo cuasi–paralelo.

creación de un hilo (para empezar)

- Se crea un hilo con

  `Thread worker = new Thread()`
- Después se inicializa el hilo y se define su comportamiento.
- Se lanza el hilo con

  `worker.start()`
- Pero en esta versión simple no hace nada. Hace falta sobreescribir el método `run()` especificando algún código útil.

la interfaz `Runnable`

- A veces no es conveniente extender la clase `Thread` porque se pierde la posibilidad de extender otro objeto.
- Es una de las razones por que existe la interfaz `Runnable` que declara nada más que el método `public void run()` y que se puede usar fácilmente para crear hilos trabajadores.

pingPONG I

```
class RunPingPONG implements Runnable {
  private String word;
  private int delay;

  RunPingPONG(String whatToSay, int delayTime) {
    word =whatToSay;
    delay=delayTime;
  }
```


pingPONG II

```
public void run() {
  try {
    for(;;) {
      System.out.print(word+" ");
      Thread.sleep(delay);
    }
  }
  catch(InterruptedException e) {
    return;
  }
}
```

pingPONG III

```
  public static void main(String[] args) {
    Runnable ping = new RunPingPONG("ping", 40);
    Runnable PONG = new RunPingPONG("PONG", 50);
    new Thread(ping).start();
    new Thread(PONG).start();
  }
}
```


construcción de `Runnable`s

Existen cinco constructores para crear hilos usando la interfaz `Runnable`.

- `public Thread(Runnable target)`
  así lo usamos en el ejemplo arriba, se pasa solamente la implementación de la interfaz `Runnable`
- `public Thread(Runnable target, String name)`
  se pasa adicionalmente un nombre para el hilo
- `public Thread(ThreadGroup group, Runnable target)`
  construye un hilo dentro de un grupo de hilos
- `public Thread(ThreadGroup group, Runnable target, String name)`
  construye un hilo con nombre dentro de un grupo de hilos
- `public Thread(ThreadGroup group, Runnable target, String name, long stackSize)`
  construye un hilo con nombre dentro de un grupo de hilos

implementación de `Runnable`

- La interfaz `Runnable` exige solamente el método `run()`, sin embargo, normalmente se implementan más métodos para crear un servicio completo que este hilo debe cumplir.
- Aunque no hemos guardado las referencias de los hilos en unas variables, los hilos *no caen* en las manos del recolector de memoria: siempre se mantiene su referencia como nodo *raíz* en el recolector de memoria.
  (incluso poner la referencia a `null` no termina el hilo)
- Para que caiga en manos del recolector tiene que haber terminado su método `run()` y no depender de un punto de sincronización.
- No se puede relanzar el mismo hilo otra vez.
- El método `run()` es público pero en muchos casos—implementando algún tipo de servicio—no se quiere dar permiso a otros ejecutar directamente el método `run()`. Para evitar eso se puede recurrir a la siguiente construcción:

`run()` no-público

```
class Service {
  private Queue requests = new Queue();
  public Service() {
    Runnable service = new Runnable() {
      public void run() {
        for(;;) realService((Job)requests.take());
      }
    };
    new Thread(service).start();
  }
  public void AddJob(Job job) {
    requests.add(job);
  }
  private void realService(Job job) {
    // do the real work
  }
}
```

explicación del ejemplo

- Crear el servicio con `Service()` lanza un nuevo hilo que actua sobre una cola para realizar su trabajo con cada tarea que encuentra ahí.
- El trabajo por hacer se encuentra en el método privado `realService()`.
- Una nueva tarea se puede añadir a la cola con `AddJob(...)`.
- **Nota:** la construcción arriba usa el concepto de clases anónimas de Java, es decir, sabiendo que no se va a usar la clase en otro sitio que no sea que en su punto de construcción, se declara directamente donde se usa.

sincronización

- A veces se quiere que un hilo actua sin que otros interfieran en la tarea.
- Es decir, se quiere una ejecución con exclusión mutua del código.
- A nivel bajo dicho concepto se llama atomicidad de las operaciones.
- (Existe otro concepto de conseguir algo parecido que veremos más adelante.)

`sinchronized`

- En Java es posible forzar la ejecución del código en un bloque de modo sincronizado, es decir, como mucho un hilo que tenga acceso exlusivo a `obj` puede ejecutar el código dentro de dicho bloque.

```
synchronized (obj) { ... }
```

- La expresión entre paréntesis `obj` tiene que evaluar a una referencia a un objeto o a un vector.
- Declarando un método con el modificador `synchronized` garantiza que dicho método se ejecuta por un sólo hilo (y ningún otro método sincronizado del mismo objeto tampoco).
- La máquina virtual instala un cerrojo (mejor dicho, un monitor, ya veremos dicho concepto más adelante) que se cierra de forma atómica antes de entrar en la región crítica y que se abre antes de salir.

métodos syncronizados

- Declarar un método como

```
synchronized void f() { ... }
```

  es equivalente a usar un bloque sincronizado en su interior:

```
void f() { synchronized(this) { ... } }
```

- Los monitores (implementados en la MVJ) permiten que el mismo hilo puede acceder a otros métodos o bloques sincronizados del mismo objeto sin problema.
- Se libera el cerrojo sea el modo que sea que termine el método.
- Los constructores no se pueden declarar `synchronized`.

syncronización y herencia

- No hace falta mantener el modo sincronizado sobreescribiendo métodos síncronos mientras se extiende una clase. (No se puede *forzar* un método sincronizada en una interfaz.)
- Sin embargo, una llamada al método de la clase súperior (con `super.`) sigue funcionando de modo síncrono.
- Los métodos estáticos también se pueden declarar `synchronized` garantizando su ejecución de manera exclusiva entre varios hilos.

protección de miembros estáticos

En ciertos casos se tiene que proteger el acceso a miembros estáticos con un cerrojo. Para conseguir eso es posible sincronizar con un cerrojo de la clase, por ejemplo:

```
class MyClass {
  static private int nextID;
  ...
  MyClass() {
    synchronized(MyClass.class) {
      idNum=nextID++;
    }
  }
  ...
}
```


¡Ojo con el concepto!

Declarar un bloque o un método como síncrono solo prevee que ningún otro hilo pueda ejecutar al mismo tiempo dicha región crítica (o otra sincronizada con el mismo objeto), sin embargo, cualquier otro código asíncrono puede ser ejecutado mientras tanto y su acceso a variables críticas puede dar como resultado fallos en el programa.

objetos síncronos

Se obtienen objetos totalmente sincronizados siguiendo las reglas:

- todos los métodos son `synchronized`,
- no hay miembros/atributos públicos,
- todos los métodos son `final`,
- se inicializa siempre todo bien,
- el estado del objeto se mantiene siempre consistente incluyendo los casos de excepciones.

páginas del manual

Se recomienda estudiar detenidamente las páginas del manual de Java que estén relacionados con el concepto de hilo (mira también las referencias en el boletín de prácticas).

atomicidad en Java

- Solo las asignaciones a variables de tipos simples de 32 bits son atómicas.
- `long` y `double` no son simples en este contexto porque son de 64 bits, hay que declararlas `volatile` para obtener acceso atómico.

limitaciones para la programación concurrente

- no se puede interrumpir la espera a un cerrojo (una vez llegado a un `synchronized` no hay vuelta atrás)
- no se puede influir mucho en la política del cerrojo (distinguir entre lectores y escritores, diferentes justicias, etc.)
- no se puede confinar el uso de los cerrojos (en cualquier línea se puede escribir un bloque sincronizado de cualquier objeto)
- no se puede adquirir/liberar un cerrojo en diferentes sitios, se está obligado a una estructura de bloques

paquete especial para la programación concurrente

- Por eso, y otros motivos, se ha introducido desde Java 5 un paquete especial para la programación concurrente.

  `java.util.concurrent`
- Hay que leer/estudiar todo su manual.
- Partes mencionaremos en clase en su momento.

resumen respecto a concurrencia

- `transient`
- `volatile`
- `synchronized`
- `try-catch-finally`
- `finalize`
- `Thread, Runnable`
- `java.util.concurrent`
- `java.util.concurrent.atomic`
- `wait, notity, notifyAll`
- `join, isAlive, activeCount`
- etc.

algoritmo secuencial

- Asumimos que tengamos solamente las operaciones aritméticas *sumar* y *subtraer* disponibles en un procesador fictício y queremos multiplicar dos números positivos.
- Un posible algoritmo sequencial que multiplica el número `p` con el número `q` produciendo el resultado `r` es:

```
Initially:  set p and q to positive numbers

a: set r to 0
b: loop
c:   if q equal 0 exitloop
d:   set r to r+p
e:   set q to q-1
f: endloop
g: ...
```


¿Cómo se comprueba si el algoritmo es correcto?

- Primero tenemos que decir que significa correcto.
- El algoritmo (secuencial) es correcto si
  - una vez se llega a la instrucción `g:` el valor de la variable `r` contiene el producto de los valores de las variables `p` y `q` (se refiere a sus valores que han llegado a la instrucción `a:`)
  - se llega a la instrucción `g:` en algún momento
  - y la entrada habia sido la correcta.

¿Cómo se comprueba si el algoritmo es correcto?

- Tenemos que saber que las instrucciones atómicas son correctas,
- es decir, sabemos exactamente su significado, incluyendo todos los efectos segundarios posibles.
- Luego usamos el concepto de inducción para comprobar el bucle.
- Sean p_i, q_i, y r_i los contenidos de los registros `p`, `q`, y `r`, respectivamente.
- Una invariante cuya corrección hay que comprobar con el concepto de inducción es entonces:

$$r_i + p_i \cdot q_i = p \cdot q$$

- ¿Cómo encontrar una invariante adecuada? usar *ingenio*... (RAE: Facultad del ser humano para discurrir o inventar con prontitud y facilidad.)

algoritmo concurrente

Reescribimos el algoritmo secuencial para que “funcione” con dos procesos:

```
Initially:  set p and q to positive numbers

a: set r to 0
 P0                         P1
b: loop                       loop
c:   if q equal 0 exit          if q equal 0 exit
d:   set r to r+p               set r to r+p
e:   set q to q-1               set q to q-1
f: endloop                    endloop
g: ...
```


intercalaciones posibles

- El algoritmo es no-determinista,
- en el sentido que no se sabe de antemano en qué orden (en un procesador) se van a ejecutar las instrucciones,
- o más preciso, cómo se van a intercalar las instrucciones atómicas de ambos procesos.
- El no-determinismo puede provocar situaciones con errores, es decir, el fallo ocurre solamente si las instrucciones se ejecutan en un orden específico.
- **Ejemplo:** (mostrado en pizarra)

funcionamiento correcto I

Generalmente se dice que un programa es correcto, si dada una entrada, el programa produce los resultados deseados.
Más formal:

- Sea $P(x)$ una propiedad de una variable x de entrada (aquí el símbolo x refleja cualquier conjunto de variables de entradas).
- Sea $Q(x, y)$ una propiedad de una variable x de entrada y de una variable y de salida.

funcionamiento correcto II

Se define dos tipos de funcionamiento correcto de un programa:

funcionamiento correcto parcial:
: dada una entrada a, si $P(a)$ es verdadero, y si se lanza el programa con la entrada a, entonces si el programa termina habrá calculado b y $Q(a, b)$ también es verdadero.

funcionamiento correcto total:
: dado una entrada a, si $P(a)$ es verdadero, y si se lanza el programa con la entrada a, entonces el programa termina y habrá calculado b con $Q(a, b)$ siendo también verdadero.

funcionamiento correcto III

- Un ejemplo es el cálculo de la raíz cuadrada, si x es un número flotante (por ejemplo en el estándar IEEE) queremos que un programa que calcula la raíz, lo hace correctamente para todos los números $x \geq 0$.
- Para que los procesadores puedan usar una función total (que hoy día ya es parte de las instrucciones básicas de muchos procesadores), hay que incluir los casos que x es negativo; para eso el estándar usa la codificación de `nan` (*not-a-number*).
- Calcular la raíz de un número negativo (o de `nan`) resulta en `nan`.
- (Entonces para `nan` como argumento también hay que definir todas las funciones.)

funcionamiento correcto IV

Para un programa secuencial existe solamente un orden total de las instrucciones atómicas (en el sentido que un procesador secuencial siempre sigue el mismo orden de las instrucciones... bueno, es mentira...), mientras que para un programa concurrente puedan existir varios órdenes.
Por eso se tiene que exigir:

funcionamiento correcto concurrente:
: un programa concurrente funciona correctamente, si el resultado $Q(x, y)$ no depende del orden de las instrucciones atómicas entre todos los órdenes posibles.

Funcionamiento correcto V

Entonces:

- Se debe asumir que los hilos pueden intercalarse en cualquier punto en cualquier momento.
- Los programas no deben estar basados en la suposición de que habrá un intercalado específico entre los hilos por parte del planificador.

funcionamiento correcto VI

- Para comprobar si un programa concurrente es *incorrecto* basta con encontrar una sola intercalación de instrucciones que nos lleva en un fallo.
- Para comprobar si un programa concurrente es *correcto* hay que comprobar que no se produce ningún fallo en ninguna de las intercalaciones posibles.

impracticabilidad de comprobación exhaustiva

- El número de posibles intercalaciones de los procesos en un programa concurrente crece exponencialmente con el número de unidades que maneja el planificador.
- Por eso es prácticamente imposible comprobar con la mera enumeración si un programa concurrente es correcto bajo todas las ejecuciones posibles.
- En la argumentación hasta ahora era muy importante que las instrucciones se ejecutaran de forma atómica, es decir, sin interrupción ninguna.
- Por ejemplo, se observa una gran diferencia si el procesador trabaja directamente en memoria o si trabaja con registros.

dependencia de atomicidad I

Si `increment` es atómico:

```
P1:  inc N
P2:  inc N

P2:  inc N
P1:  inc N
```

Se observa: las dos intercalaciones posibles producen el resultado correcto.

dependencia de atomicidad II

Si `increment` no es atómico:

```
P1:  load  R1,N
P2:  load  R2,N
P1:  inc   R1
P2:  inc   R2
P1:  store R1,N
P2:  store R2,N
```

Es decir, existe una intercalación que produce un resultado falso.
Ejemplos de Java:

- accesos a variables con más de 4 bytes no son atómicos.
- el operador ++ no es atómico.

¿Y?

¿El algoritmo concurrente de multiplicación con dos hilos de arriba, es correcto parcial? es correcto total?
¿Cómo compruebas el uno o el otro caso?
¿Te ocurre un algoritmo concurrente correcto?

un programa concurrente

- asumimos que tengamos un programa concurrente que quiere realizar acciones con recursos:
- si los recursos de los diferentes procesos son diferentes no hay problema (mira también actividad 2),
- si dos (o más procesos) quieren manipular el mismo recurso ¿Qué hacemos?

un programa concurrente

Tenemos básicamente tres opciones:

- se implementa exclusión mutua, es decir, solamente un proceso tiene acceso, los demás esperan
- se implementa comportamiento idéntico, es decir, desde el algoritmo se garantiza que todos los procesos actuan igual
- se implementa comportamiento transaccional, es decir, solo un proceso gana, lo que hacen los demás no influe en el resultado (con la opción que los demás lo noten o no respecto a su propia acción)
  (con la opción que cualquiera o uno específico gana).

¿Qué es exclusión mutua?

- Para evitar el acceso concurrente a recursos compartidos hace falta instalar un mecanismo de control
  - que permite la entrada de un proceso si el recurso está disponible y
  - que prohibe la entrada de un proceso si el recurso está ocupado.
- Es importante entender cómo se implementan los protocolos de entrada y salida para realizar la exclusión mutua.
- Obviamente no se puede implementar exclusión mutua usando exclusión mutua: se necesita algo más básico.
- Un método es usar un tipo de protocolo de comunicación basado en las instrucciones básicas disponibles.

estructura general basada en protocolos

Entonces el protocolo para cada uno de los participantes refleja una estructura como sigue (si protegemos código):

```
P0                        ... Pi
...                           ...
entrance protocol             entrance protocol
critical section              critical section
exit protocol                 exit protocol
...                           ...
```

instrucciones básicas

- obviamente tenemos que asumir que ciertas acciones de un proceso se puede realizar correctamente independientemente de las acciones de los demás procesos
- dichas acciones se llaman atómicas (porque son indivisibles) y se garantizan por hardware
- asumimos que podemos acceder a variables de cierto tipo (p.ej. enteros) de forma atómica con lectura y escritura (`load` y `store`)

Un posible protocolo (asimétrico)

```
  P0                                P1
a: loop                              loop
b:  non-critical section              non-critical section
c:  set v0 to true                    set v1 to true
d:  wait until v1 equals false        while v0 equals true
e:                                     set v1 to false
f:                                     wait until v0 equals false
g:                                     set v1 to true
h:  critical section                  critical section
i:  set v0 to false                   set v1 to false
j: endloop                           endloop
```


principio de la bandera

- Si ambos procesos primero levantan sus banderas
- y después miran al otro lado
- por lo menos un proceso ve la bandera del otro levantado.

comprobación con contradicción

- asumimos P0 era el último en mirar
- entonces la bandera de P0 está levantada
- asumimos que P0 no ha visto la bandera de P1
- entonces P1 ha levantado la bandera después de la mirada de P0
- pero P1 mira después de haber levantado la bandera
- entonces P0 no era el último en mirar

propiedades de interés del protocolo

Normalmente, un protocolo de entrada y salida debe cumplir con las siguientes condiciones:

- sólo un proceso debe obtener acceso a la sección crítica (garantía del acceso con exclusión mutua)
- por lo menos un proceso debe obtener acceso a la sección crítica después de un tiempo de espera *finito*.
- Obviamente se asume que ningún proceso ocupa la sección crítica durante un tiempo infinito.

propiedades de interés del protocolo

La propiedad de espera finita se puede analizar según los siguientes criterios:

justicia:
: hasta que medida influyen las peticiones de los demás procesos en el tiempo de espera de un proceso

espera:
: hasta que medida influyen los protocolos de los demás procesos en el tiempo de espera de un proceso

tolerancia a fallos:
: hasta que medida influyen posibles errores de los demás procesos en el tiempo de espera de un proceso.

análisis del protocolo asimétrico

Analizamos el protocolo de antes respecto a dichos criterios:

- ¿Está garantizado la exclusión mutua?
- ¿Influye el estado de uno (sin acceso) en el acceso del otro?
- ¿Quién gana en caso de peticiones simultaneas?
- ¿Qué pasa en caso de error?

soporte hardware

- Dependiendo de las capacidades del hardware la implementación de los protocolos de entrada y salida es más fácil o más difícil, además las soluciones pueden ser más o menos eficientes.
- Vimos y veremos que se pueden realizar protocolos seguros solamente con las instrucciones `load` y `store` de un procesador.
- Las soluciones no suelen ser muy eficientes, especialmente si muchos procesos compiten por la sección crítica. *Pero: su desarrollo y la presentación de la solución ayuda en entender el problema principal.*
- A veces no hay otra opción disponible.
- Todos los microprocesadores modernos proporcionan instrucciones básicas que permiten realizar los protocolos de forma más directo y en muchas ocaciones más eficiente.

primer intento

Usamos una variable `v` que nos indicará cual de los dos procesos tiene su turno.

```
  P0                          P1
a: loop                       loop
b:  wait until v equals P0     wait until v equals P1
c:  critical section           critical section
d:  set v to P1                set v to P0
e:  non-critical section       non-critical section
f: endloop                    endloop
```


primer intento: propiedades

- Está garantizada la exlusión mutua porque un proceso llega a su línea `c:` solamente si el valor de `v` corresponde a su identificación (que asumimos siendo única).
- Obviamente, los procesos pueden acceder al recurso solamente alternativamente, que puede ser inconveniente porque acopla los procesos fuertemente.
- Un proceso no puede entrar más de una vez seguido en la sección crítica.
- Si un proceso termina el programa o no llega más por alguna razón a su línea `d:`, el otro proceso puede resultar bloqueado.
- La solución se puede ampliar fácilmente a más de dos procesos.

primer intento en Java (comenzar es fácil)

```java
import java.lang.Thread;

public class pingpong {
  volatile static int turn;  // It's v.

public
  static void main(String[] args) {
    Thread ping1 = new Player(1);
    Thread ping2 = new Player(2);

    ping1.start();
    ping2.start();

    turn=1;
  }
}
```


primer intento en Java (comenzar es fácil)

```java
class Player extends Thread {
private int id;
public Player(int id) { this.id=id; }
public void run() {
  while(true) {                          // a:
    while(id!=pingpong.turn);            // b:
    System.out.println("ping"+id);       // c:
    pingpong.turn=id\%2+1;               // d:
  }                                      // f:
}
```

primer intento en Java (terminar no tanto)

```
import java.lang.Thread;

public class pingpong {
  volatile static int turn;  // It's v.
  static void Wait(int us) {
    try {
      Thread.sleep(us);
    } catch(InterruptedException e) {
      System.out.println("waiting␣interrupted");
    }
  }
public
  static void main(String[] args) {
    ...
  }
}
```


primer intento en Java (terminar no tanto)

```
public static void main(String[] args) {
  Thread ping1 = new Player(1);
  Thread ping2 = new Player(2);
  ping1.start();
  ping2.start();
  System.out.println("playing␣some␣seconds");
  turn=1;
  Wait(2000);
  System.out.println("waiting␣for␣players");
  turn=3;  // Try to stop :-)
  try { ping1.join(); ping2.join();
  } catch(InterruptedException e) {
    System.out.println("got␣interrupted");
  }
  System.out.println("finished");
}
```


primer intento en Java (terminar no tanto)

```
class Player extends Thread {
private int id;
public Player(int id) { this.id=id; }
public void run() {
  while(pingpong.turn!=3) {
    while(id!=pingpong.turn && pingpong.turn!=3);
    System.out.println("ping"+id);
    if(pingpong.turn==id) {
      pingpong.turn=id\%2+1;
    }
  }
}
```


primer intento en Java (terminar no tanto)

```
class Player extends Thread {
private int id;
public Player(int id) { this.id=id; }
public void run() {
  while(pingpong.turn!=3) {
    while(id!=pingpong.turn && pingpong.turn!=3);
    System.out.println("ping"+id);
    if(pingpong.turn==id) {
      pingpong.Wait(100);      // And blocking!!!
      pingpong.turn=id\%2+1;
    }
  }
}
```

segundo intento

Intentamos evitar la alternancia. Usamos para cada proceso una variable, `v0` para `P0` y `v1` para `P1` respectivamente, que indican si el correspondiente proceso está usando el recurso.

```
  P0                             P1
a: loop                          loop
b:  wait until v1 equals false    wait until v0 equals false
c:  set v0 to true                set v1 to true
d:  critical section              critical section
e:  set v0 to false               set v1 to false
f:  non-critical section          non-critical section
g: endloop                       endloop
```


segundo intento: propiedades

- Ya no existe la situación de la alternancia.
- Sin embargo: el algoritmo no está seguro, porque los dos procesos pueden alcanzar sus secciones críticas simultáneamente.
- El problema está escondido en el uso de las variables de control. `v0` se debe cambiar a verdadero solamente si `v1` sigue siendo falso.
- ¿Cuál es la intercalación maligna?

tercer intento

Cambiamos el lugar donde se modifica la variable de control:

```
  P0                             P1
a: loop                          loop
b:  set v0 to true                set v1 to true
c:  wait until v1 equals false    wait until v0 equals false
d:  critical section              critical section
e:  set v0 to false               set v1 to false
f:  non-critical section          non-critical section
g: endloop                       endloop
```


tercer intento: propiedades

- Está garantizado que ambos procesos no entren al mismo tiempo en sus secciones críticas.
- Pero se bloquean mutuamente en caso que lo intenten simultaneamente que resultaría en una espera infinita.
- ¿Cuál es la intercalación maligna?

cuarto intento

Modificamos la instrucción `c:` para dar la oportunidad que el otro proceso encuentre su variable a favor.

```
  P0                         P1
a: loop                      loop
b:  set v0 to true            set v1 to true
c:  repeat                    repeat
d:   set v0 to false           set v1 to false
e:   set v0 to true            set v1 to true
f:  until v1 equals false     until v0 equals false
g:  critical section          critical section
h:  set v0 to false           set v1 to false
i:  non-critical section      non-critical section
j: endloop                   endloop
```


cuarto intento: propiedades

- Está garantizado la exclusión mutua.
- Se puede producir una variante de bloqueo: los procesos hacen algo pero no llegan a hacer algo útil (*livelock*)
- ¿Cuál es la intercalación maligna?

algoritmo de Dekker: quinto intento

```
Initially:  v0,v1 are equal to false, v is equal to P0 o P1

  P0                          P1
a: loop                        loop
b:  set v0 to true              set v1 to true
c:  loop                        loop
d:   if v1 equals false exit     if v0 equals false exit
e:   if v equals P1              if v equals P0
f:    set v0 to false             set v1 to false
g:    wait until v equals P0      wait until v equals P1
h:    set v0 to true              set v1 to true
i:   fi                          fi
j:  endloop                     endloop
k:  critical section            critical section
l:  set v0 to false             set v1 to false
m:  set v to P1                 set v to P0
n:  non-critical section        non-critical section
o: endloop                     endloop
```


quinto intento: propiedades

El algoritmo de Dekker resuelve el problema de exclusión mutua en el caso de dos procesos, donde se asume que la lectura y la escritura de un valor íntegro de un registro se puede realizar de forma atómica.

algoritmo de Peterson

```
  P0                        P1
a: loop                    loop
b:  set v0 to true          set v1 to true
c:  set v to P0             set v to P1
d:  wait while              wait while
e:   v1 equals true          v0 equals true
f:   and v equals P0         and v equals P1
g:  critical section        critical section
h:  set v0 to false         set v1 to false
i:  non-critical section    non-critical section
j: endloop                 endloop
```


algoritmo de Lamport

o algoritmo de la panadería:

- cada proceso tira un ticket (que están ordenados en orden ascendente)
- cada proceso espera hasta que su valor del ticket sea el mínimo entre todos los procesos esperando
- el proceso con el valor mínimo accede la sección crítica

algoritmo de Lamport: observaciones

- ya se necesita un cerrojo (acceso con exclusión mutua) para acceder a los tickets,
- el número de tickets no tiene límite,
- los procesos tienen que comprobar continuadamente todos los tickets de todos los demás procesos.
- El algoritmo no es verdaderamente practicable dado que se necesita un número infinito de tickets y un número elevado de comprobaciones.
- Si se sabe el número máximo de participantes basta con un número fijo de tickets.

otros algoritmos

- Como vimos, el algoritmo de Lamport (algoritmo de la panadería) necesita muchas comparaciones de los tickets para *n* procesos.
- Existe una versión de Peterson que usa solamente variables confinadas a cuatro valores.
- Existe una generalización del algoritmo de Peterson para *n* procesos (*filter algorithm*).
- Se puede evitar la necesidad de un número infinito de tickets, si se conoce antemano el número máximo de participantes (uso de grafos de precedencia).
- Otra posibilidad es al algoritmo de Eisenberg–McGuire (que garantiza una espera mínima para *n* procesos).

límites

- Se puede comprobar que se necesita por lo menos *n* campos en la memoria común para implementar un algoritmo (con `load` and `store`) que garantiza la exclusión mutua entre *n* procesos.

Operaciones en la memoria

- Si existen instrucciones más potentes (que los simples `load` y `store`) en el microprocesador se puede realizar la exclusión mutua más fácil.
- Hoy casi todos los procesadores implementan un tipo de instrucción atómica que realiza algún cambio en la memoria al mismo tiempo que devuelve el contenido anterior de la memoria.

TAS

La instrucción `test-and-set` (TAS) implementa

- una comprobación a cero del contenido de una variable en la memoria
- al mismo tiempo que varía su contenido
- en caso que la comprobación se realizó con el resultado verdadero.

TAS

```
Initially:  vi is equal false
            C  is equal true

a: loop
b:   non-critical section
c:   loop
d:     if C equals true              ; atomic
         set C to false and exit
e:   endloop
f:   set vi to true
g:   critical section
h:   set vi to false
i:   set C to true
j: endloop
```

TAS: propiedades

- En caso de un sistema multi-procesador hay que tener cuidado que la operación `test-and-set` esté realizada en la memoria compartida.
- Teniendo solamente una variable para la sincronización de varios procesos el algoritmo arriba no garantiza una espera limitada de todos los procesos participando.
  ¿Por qué?
- ¿Cómo se puede garantizar una espera limitada?
- Para conseguir una espera limitada se implementa un protocolo de paso de tal manera que un proceso saliendo de su sección crítica da de forma explícita paso a un proceso esperando (en caso que tal proceso exista).

EXCH

La instrucción `exchange` (a veces llamado `read-modify-write`)

- intercambia un registro del procesador
- con el contenido de una dirección de la memoria
- en una instrucción atómica.

EXCH

```
Initially:  vi is equal false
            C  is equal true

a: loop
b:   non-critical section
c:   loop
d:     exchange C and vi       ; atomic exchange
e:     if vi equals true exit
f:   endloop
g:   critical section
h:   exchange C and vi
i: endloop
```


EXCH: propiedades

- Se observa lo mismo que en el caso anterior, no se garantiza una espera limitada.
- ¿Cómo se consigue?

F&A

La instrucción `fetch-and-increment` o `fetch-and-add`

- aumenta el valor de una variable en la memoria
- y devuelve el resultado
- en una instrucción atómica.

- Con dicha instrucción se puede realizar los protocolos de entrada y salida.
- ¿Cómo?
- También existe en la versión `fetch-and-add` que en vez de incrementar suma un valor dado de forma atómica.

CAS

- La instrucción `compare-and-swap` (CAS) es una generalización de la instrucción `test-and-set`.
- La instrucción trabaja con dos variables, les llamamos `C` (de *compare*) y `S` (de *swap*).
- Se intercambia el valor en la memoria por `S` si el valor en la memoria es igual que `C`.
- Es la operación que se usa por ejemplo en los procesadores de Intel y es la base para facilitar la concurrencia en la máquina virtual de Java 1.5 para dicha familia de procesadores.
- Con CAS se pueden realizar los protocolos de entrada y salida. ¿Cómo?

double CAS

Existen también unas mejoras del `CAS`, llamado *double-compare-and-swap* DCAS (Motorola), que realiza dos CAS normales a la vez, o *double-wide compare-and-swap* (Intel/AMD x86), que opera con dos punteros a la vez para el intercambio, o *single-compare double-swap* (Intel itanium), que compara un valor (puntero) pero escribe dos punteros en memoria adyacente.
El código, expresado a alto nivel, para DCAS sería:

```
if  C1 equal to V1 and C2 equal to V2
  then
    swap S1 and V1
    swap S2 and V2
    return true
  else
    return false
```


Uso de memoria común

- como hemos visto todos los protocolos necesitan variables con acceso atómico en memoría común.
- Tal hecho puede resultar en pérdidas de rendimiento, si existe una jerarquía de memoria con diferentes niveles de cachés.
- Muchos compiladores ofrecen acceso cuasi directo a las instrucciones de nivel bajo, por ejemplo la gama de compiladores GCC con sus *atomic builtins*, (`https://gcc.gnu.org/onlinedocs/gcc-4.1.2/gcc/Atomic-Builtins.html`)
  Ojo con código que tiene que ser compatible entre arquitecturas.

ABA problem

no se transmite information

- los protocolos de entrada y salida como implementados hasta ahora no transmiten información de un hilo al otro
- en el sentido que si un hilo entra en su sección crítica dos veces
- no puede averiguar si otro hilo ha (o otros hilos han) entrado mientras tanto
- este problema se concoce como ABA-problema (la secuencia *primero A, luego B, después A*, no se puede distinguir del simple hecho *solo A*)
- el problema es, por ejemplo, relevante si se compara solo punteros o referencias para averiguar posibles cambios de estado
- (lo veremos más adelante otra vez con sus posibles implicaciones)

conceptos

- El concepto de usar estructuras de datos a nivel alto libera al/a programador/a de los detalles de su implementación.
- Se puede asumir que las operaciones están implementadas correctamente y se puede basar el desarrollo del programa concurrente en un funcionamiento correcto de las operaciones de los tipos de datos abstractos.
- Las implementaciones concretas de los tipos de datos abstractos recurren a las posibilidades descritas anteriormente (o algo parecido).
- Para que se puedan utilizar con provecho hay que entender en detalle las propiedades de tales estructuras de datos.

semáforo

Un semáforo es un tipo de datos abstracto que permite el uso de un recurso de manera exclusiva cuando varios procesos están competiendo y que cumple la siguiente semántica:

- El estado interno del semáforo cuenta cuantos procesos todavía pueden utilizar el recurso. Se puede realizar, por ejemplo, con un número entero que nunca debe llegar a ser negativo.
- Existen tres operaciones con un semáforo: `init()`, `wait()`, y `signal()` que realizan lo siguiente:

operaciones del semáforo I

`init()`:

- Inicializa el semáforo antes de que cualquier proceso haya ejecutado ni una operación `wait()` ni una operación `signal()` al límite de número de procesos que tengan derecho a acceder el recurso.
- Si se inicializa con 1, se ha construido un semáforo binario.
- En lenguajes orientados a objectos, la operación `init()` se suele realizar en la construcción del objeto correspondiente.

operaciones del semáforo II

`wait()`:

- Si el estado indica cero, el proceso se queda atrapado en el semáforo hasta que sea despertado por otro proceso.
- Si el estado indica que un proceso más puede acceder el recurso se decrementa el contador y la operación termina con exito.
- La operación `wait()` tiene que estár implementada como una instrucción atómica. Sin embargo, en muchas implementaciones la operación `wait()` puede ser interrumpida.
- Normalmente existe una forma de comprobar si la salida del `wait()` es debido a una interrupción o porque se ha dado acceso al semáforo.

operaciones del semáforo III

`signal()`:

- Una vez se ha terminado el uso del recurso, el proceso lo señaliza al semáforo. Si queda algún proceso bloqueado en el semáforo, uno de ellos sea despertado, sino se incrementa el contador.
- La operación `signal()` también tiene que estár implementada como instrucción atómica. En algunás implementaciones es posible comprobar si se ha despertado un proceso con exito en caso que había alguno bloqueado.
- Para despertar los procesos se pueden implementar varias formas que se destinguen en su política de justicia (p.ej. FIFO).

uso del semáforo

El acceso mutuo a secciones críticas se arregla con un semáforo que permita el acceso a un sólo proceso

```
S.init(1)

P1                             P2
a: loop                            loop
b:   S.wait()                        S.wait()
c:   critical section                critical section
d:   S.signal()                      S.signal()
e:   non-critical section            non-critical section
f: endloop                         endloop
```


observamos los siguientes detalles

- Si algún proceso no libera el semáforo, se puede provocar un bloqueo.
- No hace falta que un proceso libere su propio recurso, es decir, la operación `signal()` puede ser ejecutada por otro proceso.
- Con semáforos simple (o binarios) no se puede imponer un orden a los procesos accediendo a diferentes recursos.

semáforos binarios/generales

Si existen en un entorno solamente semáforos binarios, se puede simular un semáforo general usando dos semáforos binarios y un contador, por ejemplo, con las variables `delay`, `mutex` y `count`.

- La operación `init()` inicializa el contador al número máximo permitido.
- El semáforo `mutex` asegura acceso mutuamente exclusivo al contador.
- El semáforo `delay` atrapa a los procesos que superan el número máximo permitido.

detalles de la implementación I

La operación `wait()` se implementa de la siguiente manera:

```
delay.wait()
mutex.wait()
decrement count
if count greater 0 then delay.signal()
mutex.signal()
```


detalles de la implementación II

La operación `signal()` se implementa de la siguiente manera:

```
mutex.wait()
increment count
if count equal 1 then delay.signal()
mutex.signal()
```


principales desventajas de semáforos

- No se puede imponer el uso correcto de las llamadas a los `wait()`s y `signal()`s.
- No existe una asociación entre el semáforo y el recurso.
- Entre `wait()` y `signal()` el usuario puede realizar cualquier operación con el recurso.

región crítica

- Un lenguaje de programación puede realizar directamente una implementación de una región crítica.
- Así parte de la responsabilidad se traslada desde el programador al compilador.
- De alguna manera se identifica que algún bloque de código se debe tratar como región crítica (así funciona Java con sus bloques sincronizados):

```
V is shared variable
region V do
  code of critical region
```


observaciones I

- El compilador asegura que la variable `V` tenga un semáforo adjunto que se usa para controlar el acceso exclusivo de un solo proceso a la región crítica.
- De este modo no hace falta que el programador use directamente las operaciones `wait()` y `signal()` para controlar el acceso con el posible error de olvidarse de algún `signal()`.
- Adicionalmente es posible que dentro de la región crítica se llame a otra parte del programa que a su vez contenga una región crítica. Si ésta está controlada por la misma variable `V` el proceso obtiene automáticamente también acceso a dicha región.

observaciones II

- Las regiones críticas no son lo mismo que los semáforos, porque no se tiene acceso directo a las operaciones `init()`, `wait()` y `signal()`.
- Con semáforos se puede emular regiones críticas pero no al revés.

regiones críticas condicionales

- En muchas situaciones es conveniente controlar el acceso de varios procesos a una región crítica por una condición.
- Con las regiones críticas simples, vistas hasta ahora, no se puede realizar tal control. Hace falta otra construcción, por ejemplo:

```
V is shared variable
C is boolean expression
region V when C do
  code of critical region
```

detalles de implementación I

Las regiones críticas condicionales funcionan internamente de la siguiente manera:

- Un proceso que quiere entrar en la región crítica espera hasta que tenga permiso.
- Una vez obtenido permiso comprueba el estado de la condición, si la condición lo permite entra en la región, en caso contrario libera el cerrojo y se pone de nuevo esperando en la cola de acceso.

detalles de implementación II

- Se implementa una región crítica normalmente con dos colas diferentes.
- Una cola principal controla los procesos que quieren acceder a la región crítica, una cola de eventos controla los procesos que ya han obtenido una vez el cerrojo pero que han encontrado la condición en estado falso.
- Si un proceso sale de la región crítica todos los procesos que quedan en la cola de eventos pasan de nuevo a la cola principal porque tienen que recomprobar la condición.

detalles de implementación III

- Nota que esta técnica puede derivar en muchas comprobaciones de la condición, todos en modo exclusivo, y puede causar pérdidas de eficiencia.
- En ciertas circunstancias hace falta un control más sofisticado del acceso a la región crítica dando paso directo de un proceso a otro.

desventajas de semáforos y regiones críticas

Todas las estructuras que hemos visto hasta ahora siguen provocando ciertos problemas en ciertas situaciones para el programador:

- El control sobre los recursos está distribuido por varios puntos de un programa.
- No hay protección de las variables de control que siempre fueron variables globales.

monitor

Por eso se usa el concepto de monitores que implementan un nivel aún más alto de abstracción facilitando el acceso a recursos compartidos.
Un monitor es un tipo de datos abstracto que contiene

- un conjunto de procedimientos de control de los cuales solamente un subconjunto es visible desde fuera,
- un conjunto de datos privados, es decir, no visibles desde fuera.

detalles de implementación de un monitor

- El acceso al monitor está permitido solamente a través de los métodos públicos y el compilador garantiza exclusión mutua para todos los accesos.
- La implementación del monitor controla la exclusión mutua con colas de entrada que contengan todos los procesos bloqueados.
- Pueden existir varias colas y el controlador del monitor elige de cual cola se va a escoger el siguiente proceso para actuar sobre los datos.
- Un monitor no tiene acceso a variables exteriores con el resultado de que su comportamiento no puede depender de ellos.
- Una desventaja de los monitores es la exclusividad de su uso, es decir, la concurrencia está limitada si muchos procesos hacen uso del mismo monitor.

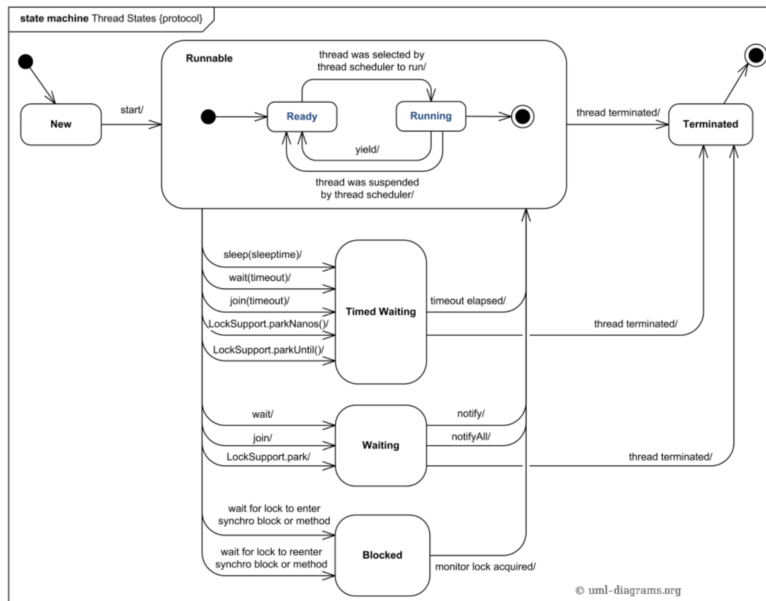
sincronización condicional

- Un aspecto que se encuentra en muchas implementaciones de monitores es la sincronización condicional, es decir, mientras un proceso está ejecutando un procedimiento del monitor (con exclusión mutua) puede dar paso a otro proceso liberando el cerrojo.
- Estas operaciones se suele llamar `wait()` o `delay()`. El proceso que ha liberado el cerrojo se queda bloqueado hasta que otro proceso le despierta de nuevo.
- Este bloqueo temporal está realizado dentro del monitor (dicha técnica se refleja en Java con `wait()` y `notify()/notifyAll()`).
- La técnica permite la sincronización entre procesos porque actuando sobre el mismo recurso los procesos pueden cambiar el estado del recurso y pasar así información de un proceso al otro.

disponibilidad de monitores

- Lenguajes de alto nivel que facilitan la programación concurrente suelen tener monitores implementados dentro del lenguaje (por ejemplo en Java).
- El uso de monitores es bastante costoso, porque se puede perder eficiencia por bloquear los procesos innecesariamente y el trabajo adicional por el uso del monitor..
- Por eso se intenta aprovechar de primitivas más potentes para aliviar este problema (alternativas libres de cerrojos (*lock free*) y/o libres de espera (*wait free*)).

Java máquina de estados de hilos



desventajas de uso de sincronización a alto nivel

- No se distingue entre accesos de solo lectura y de escritura que limita la posibilidad de accesos en paralelo.
- Cualquier interrupción (p.ej. por falta de página de memoria) relantiza el avance de la aplicación,
- por eso las MVJ usan los procesos del sistema operativo para implementar los hilos, así el S.O. puede conmutar a otro hilo.
- Sigue presente el problema de llamar antes a `notify()`, o `notifyAll()` que a `wait()` (*race condition*).

alternativas

- estructuras de datos concurrentes (mira `java.util.conconcurrent`)
- uso de memoria transaccional

un caso reciente

www.flexcoin.com

Update (March 4 2014): During the investigation into stolen funds we have determined that the extent of the theft was enabled by a flaw within the front-end (???). The attacker logged into the flexcoin front end from IP address XXX under a newly created username and deposited to address XXX. The coins were then left to sit until they had reached 6 confirmations.
The attacker then successfully exploited a flaw in the code which allows transfers between flexcoin users. By sending thousands of simultaneous requests, the attacker was able to *move* coins from one user account to another until the sending account was overdrawn, before balances were updated. This was then repeated through multiple accounts, snowballing the amount, until the attacker withdrew the coins.

un caso reciente
www.flexcoin.com

Flexcoin has made every attempt (???) to keep our servers as secure as possible, including regular testing. In our approx. 3 years of existence we have successfully repelled thousands of attacks. But in the end, this was simply not enough. Having this be the demise of our small company, after the endless hours of work we've put in, was never our intent. We've failed our customers, our business, and ultimatley the Bitcoin community.

un caso reciente
www.flexcoin.com

Look and smile:

```
mybalance = database.read("account-number")
newbalance = mybalance - amount
database.write("account-number", newbalance)
dispense_cash(amount)   // or send bitcoins to customer
```

(taken from http://hackingdistributed.com/2014/04/06/another-one-bites-the-dust-flexcoin/)

problema

Un bloqueo se produce cuando un proceso está esperando algo que nunca se cumple.
Ejemplo:
Cuando dos procesos P_0 y P_1 quieren tener acceso simultaneamente a dos recursos r_0 y r_1, es posible que se produzca un bloqueo de ambos procesos. Si P_0 accede con éxito a r_1 y P_1 accede con éxito a r_0, ambos se quedan atrapados intentando tener acceso al otro recurso.

condiciones necesarias

Se tienen que cumplir cuatro condiciones para que sea posible que se produzca un bloqueo entre procesos:

1. los procesos tienen que compartir recursos con exclusión mutua
2. los procesos quieren acceder a un recurso más mientras ya tienen acceso exclusivo a otro
3. los recursos solo permiten ser usados por menos procesos que lo intentan al mismo tiempo
4. existe una cadena circular entre peticiones de procesos y asignaciónes de recursos

funcionamiento parcial

Un problema adicional con los bloqueos es que es posible que el programa siga funcionando correctamente según la definición, es decir, el resultado obtenido es el resultado deseado, pero algunos de sus procesos están bloqueados durante la ejecución (es decir, se produjo solamente un bloqueo parcial).

técnicas de actuar

Existen algunas técnicas que se pueden usar para que no se produzcan bloqueos:

- Detectar y actuar
- Evitar
- Prevenir

detectar y actuar

Se implementa un proceso adicional que vigila si los demás forman una cadena circular.
Más preciso, se define el grafo de asignación de recursos:

- Los procesos y los recursos representan los nodos de un grafo.
- Se añade cada vez una arista entre un nodo tipo recurso y un nodo tipo proceso cuando el proceso ha obtenido acceso exclusivo al recurso.
- Se añade cada vez una arista entre un nodo tipo recurso y un nodo tipo proceso cuando el proceso está pidiendo acceso exclusivo al recurso.
- Se eliminan las aristas entre proceso y recurso y al revés cuando el proceso ya no usa el recurso.

actuación

Cuando se detecta en el grafo resultante un ciclo, es decir, cuando ya no forma un grafo acíclico, se ha producido una posible situación de un bloqueo.
Se puede reaccionar de dos maneras si se ha encontrado un ciclo:

- No se da permiso al último proceso de obtener el recurso.
- Sí, se da permiso, pero una vez detectado el ciclo se aborta todos o algunos de los procesos involucrados.

desventaja

Sin embargo, las técnicas pueden dar como resultado que el programa no avance, incluso, el programa se puede quedar atrapado haciendo trabajo inútil: crear situaciones de bloqueo y abortar procesos continuamente.

evitar

El sistema no da permiso de acceso a recursos si es posible que el proceso se bloquee en el futuro.
Un método es el algoritmo del banquero (Dijkstra) que es un algoritmo centralizado y por eso posiblemente no muy practicable en muchas situaciones.
Se garantiza que todos los procesos actuan de la siguiente manera en dos fases:

1. primero se obtiene todos los cerrojos necesarios para realizar una tarea, eso se realiza solamente si se puede obtener todos a la vez,
2. después se realiza la tarea durante la cual posiblemente se liberan recursos que no son necesarias.

ejemplo

Asumimos que tengamos 3 procesos que actuan con varios recursos.
El sistema dispone de 12 recursos.

proceso	recursos pedidos	recursos reservados
A	4	1
B	6	4
C	8	5
suma	18	10

es decir, de los 12 recursos disponibles ya 10 están ocupados. La única forma que se puede proceder es dar el acceso a los restantes 2 recursos al proceso B. Cuando B haya terminado va a liberar sus 6 recursos que incluso pueden estar distribuidos entre A y C, así que ambos también pueden realizar su trabajo.
Con un argumento de inducción se verifica fácilmente que nunca se llega a ningún bloqueo.

prevenir

Se puede prevenir el bloqueo siempre y cuando se consiga que alguna de las condiciones necesarias para la existencia de un bloqueo no se produzca.

1. los procesos tienen que compartir recursos con exclusión mutua
2. los procesos quieren acceder a un recurso más mientras ya tienen acceso exclusivo a otro
3. los recursos no permiten ser usados por más de un proceso al mismo tiempo
4. existe una cadena circular entre peticiones de procesos y alocación de recursos

prevenir exclusión mutua

los procesos tienen que compartir recursos con exclusión mutua:

- No se de a un proceso directamente acceso exclusivo al recurso, si no se usa otro proceso que realiza el trabajo de todos los demás manejando una cola de tareas (por ejemplo, un demonio para imprimir con su cola de documentos por imprimir).

prevenir accesos consecutivos

los procesos quieren acceder a un recurso más mientras ya tienen acceso exclusivo a otro:

- Se exige que un proceso pida todos los recursos que va a utilizar al comienzo de su trabajo

prevenir uso único

los recursos no permiten ser usados por más de un proceso al mismo tiempo:

- Se permite que un proceso aborte a otro proceso con el fin de obtener acceso exclusivo al recurso. Hay que tener cuidado de no caer en *livelock*
- (Separar lectores y escritores alivia este problema también.)

prevenir ciclos

existe una cadena circular entre peticiones de procesos y alocación de recursos:

- Se ordenan los recursos línealmente y se fuerza a los procesos que accedan a los recursos en el orden impuesto. Así es imposible que se produzca un ciclo.

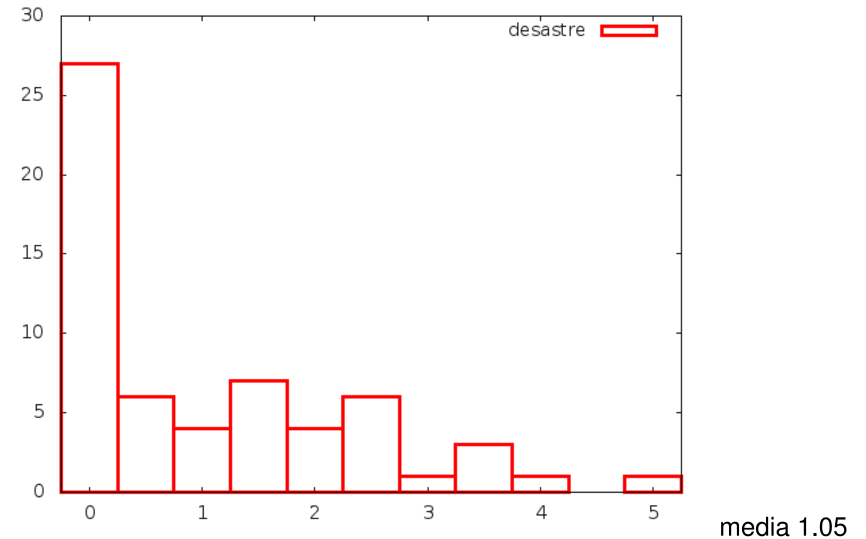
Bloqueo en Java

Un ejemplo de un bloqueo en Java muestra el siguiente trozo de código, incluso si se asume que un hilo ya está durmiendo. ¿Por qué?

```
hilo0:                           hilo1:

synchronized(A) {                synchronized(B) {
  ...                              ...
  synchronized(B) {                synchronized(A) {
    ...                              ...
    A.notify();                      B.notify();
    B.wait();                        A.wait();
  }                                }
}                                }
```


Prueba corta



media 1.05

seguridad y vivacidad/viveza

Un programa concurrente puede fallar por varias razones, las cuales se pueden clasificar entre dos grupos de propiedades:

seguridad: Esa propiedad indica que no está pasando nada malo en el programa, es decir, el programa no ejecuta instrucciones que no deba hacer (“safety property”).

vivacidad: Esa propiedad indica que está pasando continuamente algo bueno durante la ejecución, es decir, el programa consigue algún progreso en sus tareas o en algún momento en el futuro se cumple una cierta condición (“liveness property”).

propiedades de seguridad

Las propiedades de seguridad suelen ser algunas de las invariantes del programa que se tienen que introducir en las comprobaciones del funcionamiento correcto.

Corrección: El algoritmo usado es correcto.

Exclusión mutua: El acceso con exclusión mutua a regiones críticas está garantizado

Sincronización: Los procesos cumplen con las condiciones de sincronización impuestos por el algoritmo

Interbloqueo: No se produce ninguna situación en la cual todos los procesos participantes quedan atrapados en una espera a una condición que nunca se cumpla.

propiedades de vivacidad

inanición

- Un proceso puede "morirse" por inanición ("starvation"), es decir, un proceso o varios procesos siguen con su trabajo pero otros nunca avanzan por ser excluidos de la competición por los recursos (por ejemplo en Java el uso de `suspend()` y `resume()` no está recomendado por esa razón).
- Existen problemas donde la inanición no es un problema real o es muy improbable que ocurra, es decir, se puede aflojar las condiciones a los protocolos de entrada y salida.

propiedades de vivacidad

Bloqueo activo: Puede ocurrir el caso que varios procesos están continuamente competiendo por un recurso de forma activa, pero ningúno de ellos lo consigue ("livelock").

Cancelación: Un proceso puede ser terminado desde fuera sin motivo correcto, dicho hecho puede resultar en un bloqueo porque no se ha considerado la necesidad que el proceso debe realizar tareas necesarias para liberar recursos (por ejemplo, en Java el uso del `stop()` no está recomendado por esa razón).

Espera activa: Un proceso está comprobando continuamente una condición malgastando de esta manera tiempo de ejecución del procesador.

justicia entre procesos

Cuando los procesos compiten por el acceso a recursos compartidos se pueden definir varios conceptos de justicia, por ejemplo:

justicia débil: si un proceso pide acceso continuamente, le será dado en algún momento,

justicia estricta: si un proceso pide acceso infinitamente veces, le será dado en algún momento,

espera limitada: si un proceso pide acceso una vez, le será dado antes de que otro proceso lo obtenga más de una vez,

espera ordenada en tiempo: si un proceso pide acceso, le será dado antes de todos los procesos que lo hayan pedido más tarde.

comentarios

- Los dos primeros conceptos son conceptos teóricos porque dependen de términos *infinitamente* o *en algún momento*, sin embargo, pueden ser útiles en comprobaciones formales.
- En un sistema distribuido la ordenación en tiempo no es tan fácil de realizar dado que la noción de tiempo no está tan clara.
- Normalmente se quiere que todos los procesos manifesten algún progreso en su trabajo (pero en algunos casos inanición controlada puede ser tolerada).

espera activa de procesos

- El algoritmo de Dekker y sus parecidos provocan una espera activa de los procesos cuando quieren acceder a un recurso compartido. Mientras están esperando a entrar en su región crítica no hacen nada más que comprobar el estado de alguna variable.
- Normalmente no es aceptable que los procesos permanezcan en estos bucles de espera activa porque se está gastando potencia del procesador inútilmente.
- Un método mejor consiste en suspender el trabajo del proceso y reanudar el trabajo cuando la condición necesaria se haya cumplido. Naturalmente dichos técnicas de control son más complejas en su implementación que la simple espera activa.

evitar espera infinita o inanición de procesos

- Se implementa, por ejemplo, el acceso a recursos compartidos siguiendo un orden FIFO, es decir, los procesos tienen acceso en el mismo orden en que han pedido vez.
- Se asignan prioridades a los procesos de tal manera que cuanto más tiempo un proceso tiene que esperar más alto se pone su prioridad con el fin que en algún momento su prioridad sea la más alta.
- ¡Ojo! ¿Qué se hace si todos tienen la prioridad más alta?
- Existen más técnicas... ¿Cuáles?

problema clásico

El problema del productor y consumidor es un ejemplo clásico de programa concurrente y consiste en la situacion siguiente:
de una parte se produce algún producto (datos en nuestro caso) que se coloca en algún lugar (una cola en nuestro caso) para que sea consumido por otra parte. Como algoritmo obtenemos:

```
producer:                    consumer:
  forever                      forever
    produce(item)                take(item)
    place(item)                  consume(item)
```


requerimientos

Queremos garantizar que el consumidor no coja los datos más rápido de lo que los está produciendo el productor. Más concreta:

1. el productor puede generar sus datos en cualquier momento, pero no debe producir nada si no lo puede colocar
2. el consumidor puede coger un dato solamente cuando hay alguno
3. para el intercambio de datos se usa una estructura de datos compartida a la cual ambos tienen acceso,
4. si se usa una cola se garantiza un orden temporal
5. ningún dato no está consumido una vez haber sido producido (por lo menos se descarta...)

cola infinita

Si la cola puede crecer a una longitud infinita (siendo el caso cuando el consumidor consume más lento de lo que el productor produce), basta con la siguiente solucion que garantiza exclusion mutua a la cola:

```
producer:                    consumer:
  forever                      forever
    produce(item)                itemsReady.wait()
    place(item)                  take(item)
    itemsReady.signal()          consume(item)
```

donde `itemsReady` es un semáforo general que se ha inicializado al principio con el valor `0`.

correccion

El algoritmo es correcto, lo que se vee con la siguiente prueba. Asumimos que el consumidor adelanta el productor. Entonces el número de `wait()`s (terminados) tiene que ser más grande que el número de `signal()`s:

```
      #waits > #signals
==>   #signals - #waits < 0
==>   itemsReady < 0
```

y la última línea es una contradiccion a la invariante del semáforo.

más participantes

Queremos ampliar el problema introduciendo más productores y más consumidores que trabajen todos con la misma cola. Para asegurar que todos los datos estén consumidos lo más rápido posible por algún consumidor disponible tenemos que proteger el acceso a la cola con un semáforo binario (llamado `mutex` abajo):

```
producer:                    consumer:
  forever                      forever
    produce(item)                itemsReady.wait()
    mutex.wait()                 mutex.wait()
    place(item)                  take(item)
    mutex.signal()               mutex.signal()
    itemsReady.signal()          consume(item)
```


cola finita

- Normalmente no se puede permitir que la cola crezca infinitamente, es decir, hay que evitar produccion en exceso también.
- Como posible solucion introducimos otro semáforo general (llamado `spacesLeft`) que cuenta cuantos espacios quedan libres en la cola.
- Se inicializa el semáforo con la longitud máxima permitida de la cola.
- Un productor queda bloqueado si ya no hay espacio en la cola y un consumidor señala su consumision.

cola finita

```
producer:                    consumer:
  forever                      forever
    spacesLeft.wait()            itemsReady.wait()
    produce(item)                mutex.wait()
    mutex.wait()                 take(item)
    place(item)                  mutex.signal()
    mutex.signal()               consume(item)
    itemsReady.signal()          spacesLeft.signal()
```


otras estructuras compartidas

- En un sistema con multiples productores y/o consumidores, puede ser difícil establecer un orden temporal con una semántica adecuada.
- Se puede aflojar la condición de usar una cola, y usar estructuras de datos que permitan más concurrencia.
- Un ejemplo simple serían vectores de contenedores inspeccionados en orden cíclico por los productores y consumidores.

granularidad

Se suele distinguir concurrencia

- de grano fino
  es decir, se aprovecha de la ejecución de operaciones concurrentes a nivel del procesador (hardware)
- a grano grueso
  es decir, se aprovecha de la ejecución de procesos o aplicaciones a nivel del sistema operativo o a nivel de la red de ordenadores

clases de architecturas

Una clasificación clásica de ordenadores paralelos es:

- SIMD (*single instruction multiple data*)
- MISD (*multiple instruction single data*)
- MIMD (*multiple instruction multiple data*)

procesadores modernos

- Hoy día, concurrencia a grano fino es estándar en los microprocesadores.
- El la familia de los procesadores de Intel, por ejemplo, existen las instrucciones MMX, SSE, SSE2, SSE3, SSSE3 etc. que realizen según la clasificación SIMD operaciones en varios registros en paralelo.
- Ya están en el marcado los procesadores con multiples núcleos, es decir, se puede programar con varios procesadores que a su vez puedan ejecutar varios hilos independientes.

graphics processing units (GPU)

La programación paralela y concurrente (y con pipeline) se revive actualmente en la programación de las GPUs (graphics processing units) que son procesadores especializados para su uso en tarjetas gráficas que cada vez se usa más para otros fines de cálculo numérico.
Los procesadores suelen usar solamente precisión simple, pero ya hay componentes con precisión doble.

conmutación

multi–programación o *multi–programming*: los procesos se ejecutan en hardware distinto

multi–procesamiento o *multi–processing*: Se aprovecha de la posibilidad de multiplexar varios procesos en un solo procesador.

multi–tarea o *multi–tasking*: El sistema operativo (muchas veces con la ayuda de hardware específico) realiza la ejecución de varios procesos de forma cuasi–paralelo distribuyendo el tiempo disponible a las secuencias diferentes (*time–sharing system*) de diferentes usuarios (con los debidas medidas de seguridad).

computación en red

- La visión de ‘computación en la red’ no es nada más que un gran sistema MIMD.
- Existe una nueva tendencia de ofrecer en vez de aplicaciones para instalar en el cliente, una interfaz hacia un servicio (posiblemente incorporando una red privada virtual) que se ejecute en un conjunto de servidores.
- Existe una nueva tendencia de usar un llamado GRID de superordenadores para resolver problemas grandes (y distribuir el uso de los superordenadores entre más usuarios).

mecanismos de conmutación

Existen dos puntos de vista relacionados con el mecanismo de conmutación

- el mecanismo de conmutación es *independiente* del programa concurrente
  (eso suele ser el caso en sistemas operativos),
- el mecanismo de conmutación es *dependiente* del programa concurrente
  (eso suele ser el caso en sistemas en tiempo real),
- En el segundo caso es imprecindible incluir el mecanismo de conmutación en el análisis del programa.

mecanismos de conmutación

- Al desarrollar un programa concurrente, no se debe asumir ningún comportamiento específico del planificador (siendo la unidad que realiza la conmutación de los procesos).
- No obstante, un planificador puede analizar los programas concurrentes durante el tiempo de ejecución para adaptar el mecanismo de conmutación hacia un mejor rendimiento/equilibrio entre usuarios/procesos (ejemplo: automatic “nice” en un sistema Unix).
- También los sistemas suelen ofrecer unos parámetros de control para influir en las prioridades de los procesos que se usa como un dato más para la conmutación.

memoria compartida homogéneo (SMP)

- Sin una memoria compartida no existe concurrencia (se necesita por lo menos unos registros con acceso común).
- Existen varios tipos de arquitecturas de ordenadores (académicos) que fueron diseñadas especialmente para la ejecución de programas concurrentes o paralelos con una memoria compartida (por ejemplo los proyectos NYU, SB-PRAM, o Tera)
- Muchas ideas de estos proyectos se encuentran hoy día en los microprocesadores modernos, sobre todo en los protocolos que controlan la coherencia de los cachés.

memoria compartida homogéneo

- La reordenación automática de instrucciones,
- la división de instrucciones en ensemblador en varios fases de ejecución,
- la intercalación de fases de instrucciones en los procesadores
- el procesamiento en pipelines,
- la aparencia de interrupciones y excepciones,
- la jerarquia de cachés
- son propiedades desafiantes para la programación concurrente correcta con procesadores modernos y la traducción de conceptos de alto nivel del lenguaje de programación a código ejecutable no es nada fácil (y no igual en todos los entornos).

memoria compartida heterogéneo

- Sin embargo, no hace falta que se ejecute un programa en unidades similares para obtener concurrencia.
- La concurrencia está presente también en sistemas heterógenos, por ejemplo, aquellos que solapan el trabajo de entrada y salida con el resto de las tareas (discos duros).

comunicación y sincronización

La comunicación y sincronización entre procesos funciona

- mediante una memoria compartida (*shared memory*) a la cual pueden acceder todos los procesadores a la vez o
- mediante el intercambio de mensajes usando una red conectando los diferentes procesadores u ordenadores, es decir, procesamiento distribuido (*distributed processing*).
- Sin embargo, siempre hace falta algún tipo de memoria compartida para realizar la comunicación entre procesos, solamente que en algunos casos dicha memoria no es accesible en forma directa por el programador.

sistemas híbridos

También existen mezclas de todo tipo de estos conceptos, por ejemplo, un sistema que use multi–procesamiento con hilos y procesos en cada procesador de un sistema distribuido simulando una memoria compartida al nivel de la aplicación.

algunas herramientas para C/C++ (no exhaustivas)

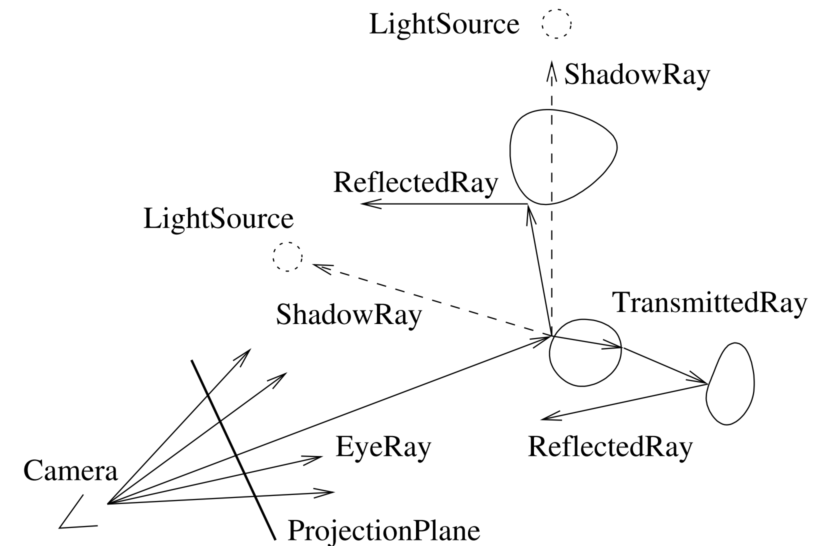
- ACE (Adaptive Communications Environment)
  `http://www.cs.wustl.edu/~schmidt/ACE.html`
  (usa patrones de diseño de alto nivel, p.ej.; proactor)
- Intel Concurrent Collections (C++, 2012, versión 0.7)
  `http://software.intel.com/en-us/articles/intel-concurrent-collections-for-cc/` (lenguaje de preprocesado para expresar concurrencia)
- Cilk/Cilk++/Cilk Plus (2011)
  `http://software.intel.com/en-us/articles/intel-cilk-plus/`
  (extensión de C/C++ para expresar concurrencia)
- Intel Thread Building Blocks (2012, version 4.0)
  `http://threadingbuildingblocks.org/`
  (C++ template librería para expresar concurrencia)

algunas herramientas para C/C++ (no exhaustivas)

- OpenMP
  http://openmp.org/wp/
  (C-preprocessor (+librería embebido) para expresar concurrencia)
- Noble (www.non-blocking.com)
- Qt threading library (http://doc.qt.nokia.com/)
- pthreads, boost::threads, Zthreads
  (uso directo de programación multi-hilo)
- próximo/ya estándar de C++ (C++11, 2011)
  (http://www.open-std.org/jtc1/sc22/wg21/)
- Usa la Red para buscar más información, aquí unos ejemplos.

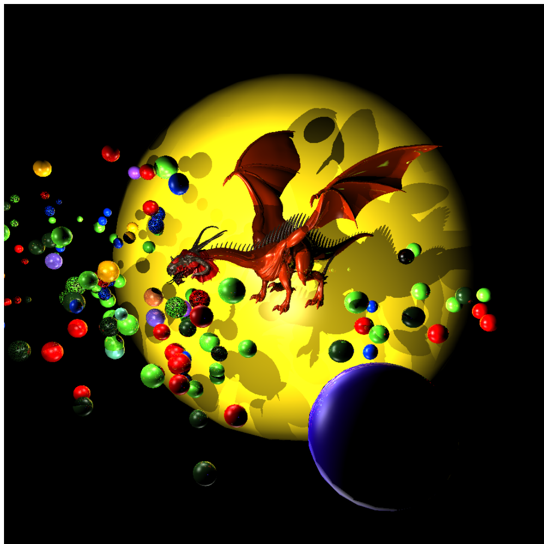
Trazado de rayos concurrente

el principio



Trazado de rayos concurrente

el resultado



Trazado de rayos concurrente

la paralelización (simple)

distribución de trabajo: bloques de píxeles de la imagen

distribución de datos:

- lectura concurrente de estructuras de datos de la escena
- escritura exclusiva de bloques en la imagen
- datos locales de los procesos incluido cachés en software (básicamente para detección de sombras y aceleración del trazado)

Trazado de rayos concurrente
la programación (ejemplo)

- programa en C++
- uso de OpenMP
- duplicación de los cachés
  (arrays con acceso mediante el identificador del hilo)
- paralelización del bucle principal

Trazado de rayos concurrente
un programa

Rayon con OpenMP

Trazado de rayos concurrente
observaciones

- con un buen diseño la paralelización es factible
- hay que estudiar bien las herramientas
- se observan las mejoras de rendimiento esperables

comentarios

- Un ingeniero informático no solo usa herramientas ya existentes, sino adapta aquellas que hay a las necesidades concretas y/o inventa nuevas herramientas para avanzar. Eso se llama innovación tecnológica.
- En el ámbito de la concurrencia son más importantes los conceptos que las tecnologías dado que las últimas están actualmente en un proceso de cambio permanente.
- Se debe comparar un programa concurrente/paralelo con el mejor programa secuencial (conocido) para evaluar el rendimiento.