Arquiteturas tolerantes a falhas

micro sistemas COTS

Taisy Silva Weber
UFRGS

Arquiteturas tolerantes a falhas

- ✓ arquitetura: nível eficaz para suportar TF
  - ✓ componentes
  - ✓ conexões

processadores, memórias, controladores, interfaces

barramentos ou linhas de comunicação

Tolerância a falhas de um sistema pode ser implementada ou por hardware, ou por software, ou ambas.

Hardware: mais eficiente.
Software: mais flexível.

Arquiteturas tolerantes a falhas

Existiram máquinas com o nome de **tolerantes a falhas**. Atualmente seriam chamadas de disponibilidade contínua ou alta disponibilidade.

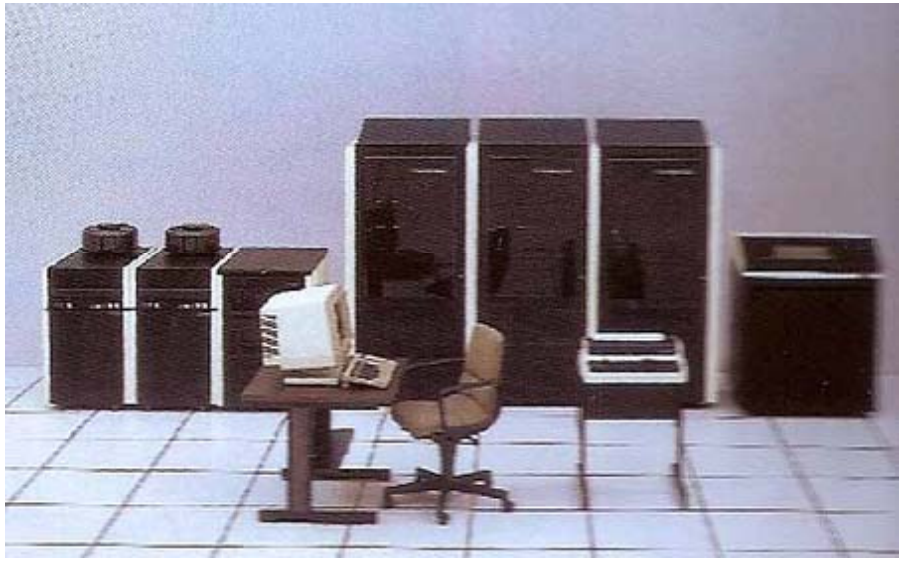
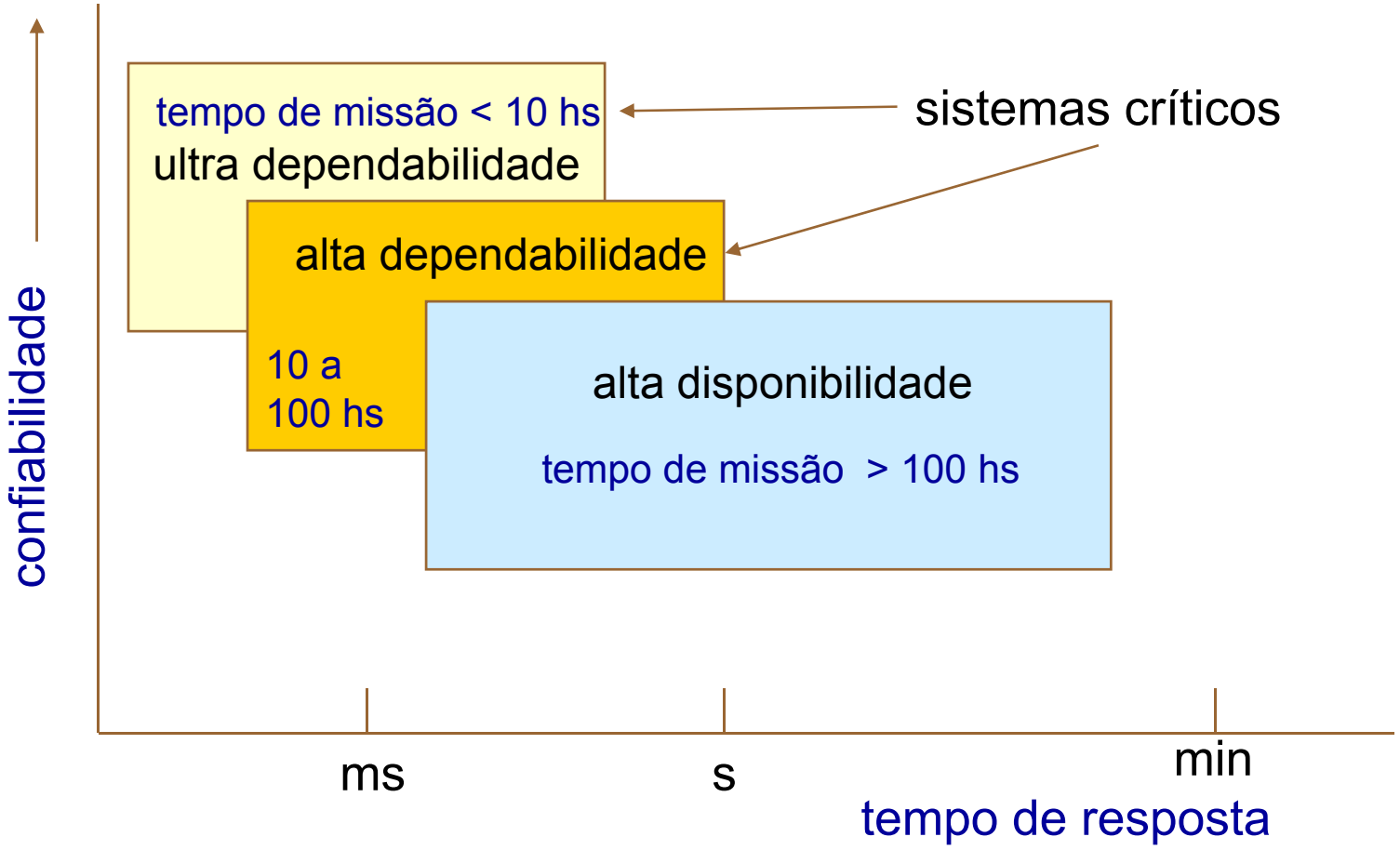


foto: http://www.siliconvalleyfamilytree.org/home/tandem_computers

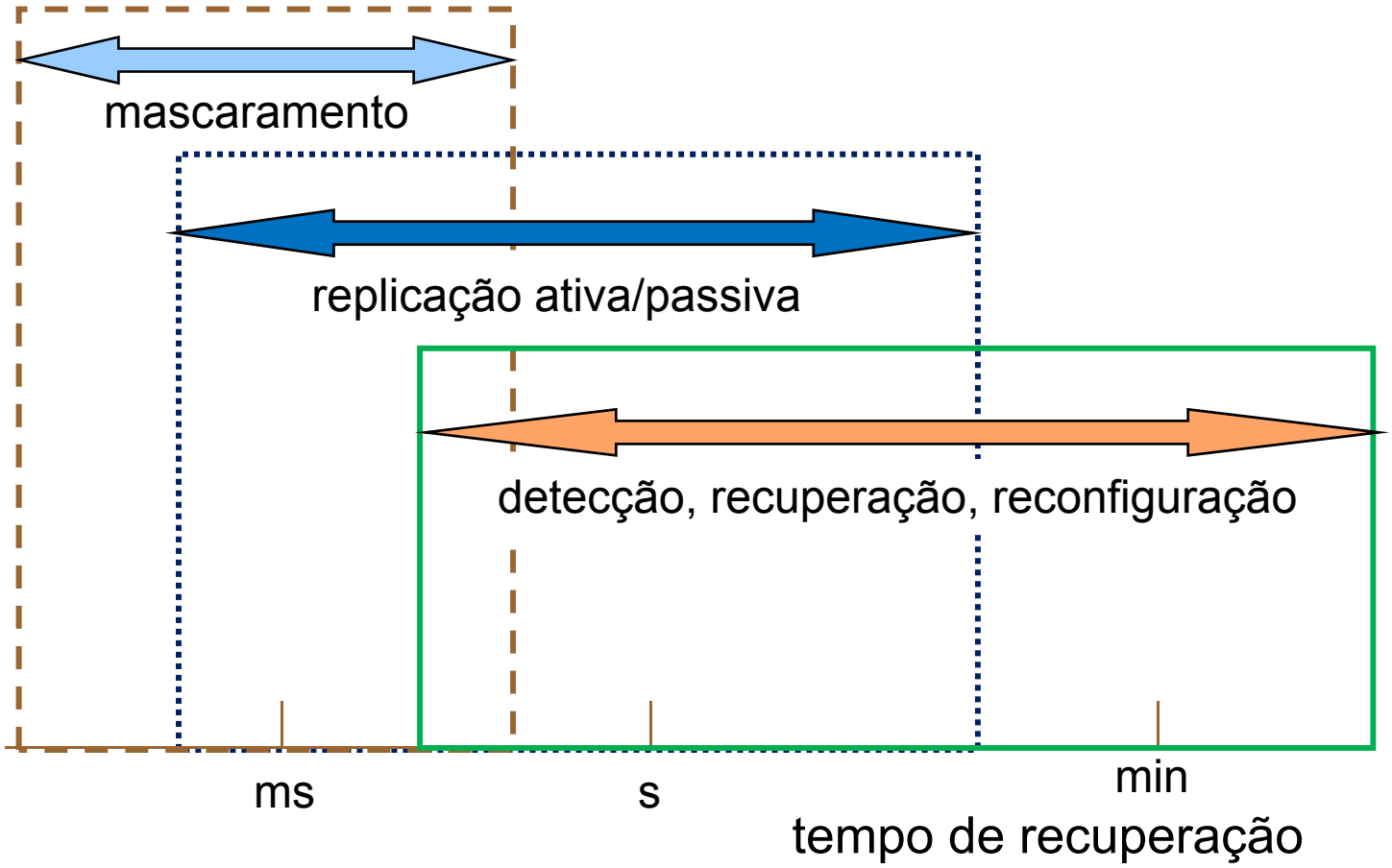
TF (aqui) é usada no sentido de arquiteturas que empregam qualquer das técnicas de tolerância a falhas.

Domínio de aplicação

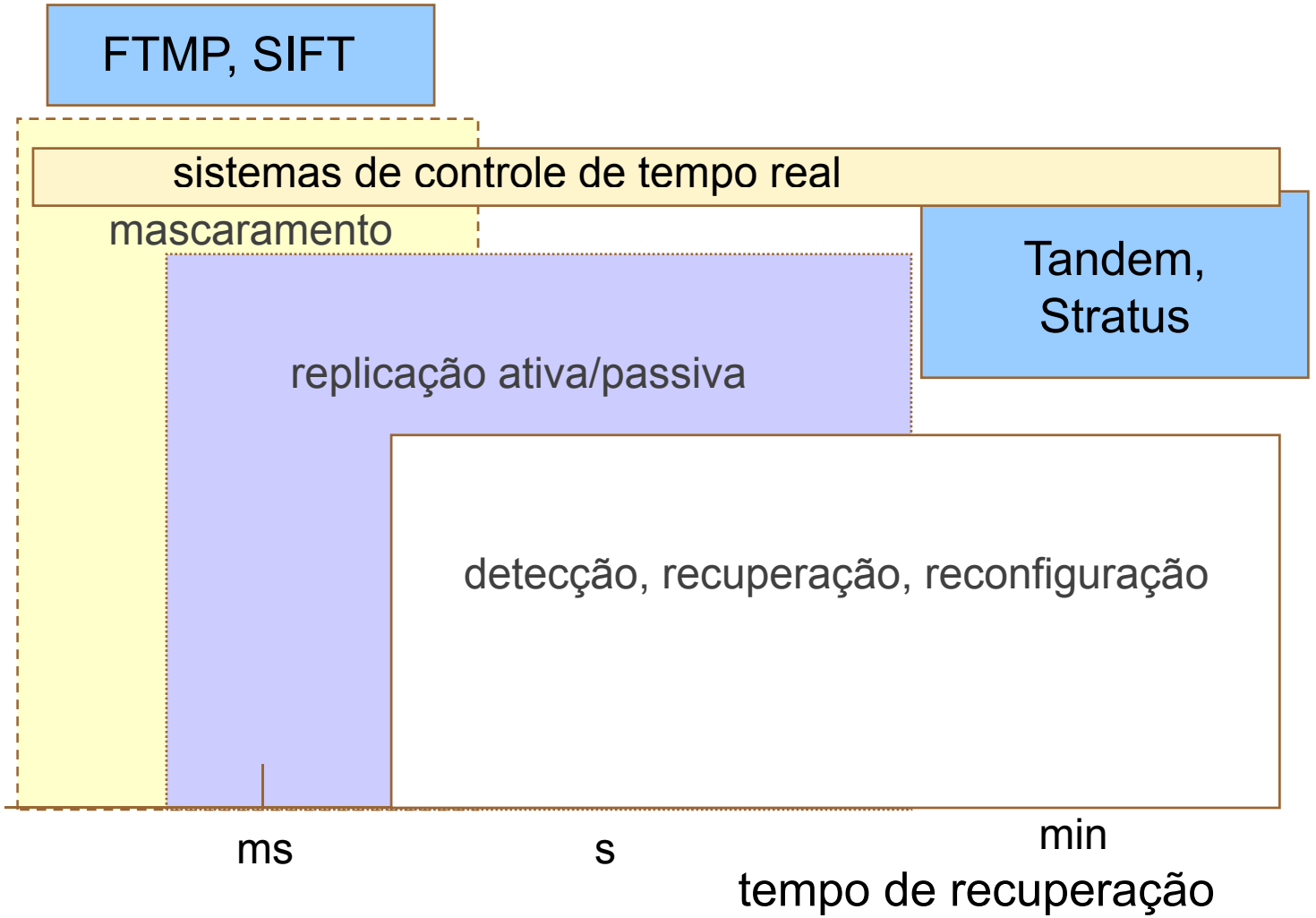
SURI, N.; WALTER, C.J.; HUGUE, M.M. **Advances in ultra-dependable distributed systems**. IEEE Computer Society Press. Los Alamitos. 1995.



Domínio de técnicas



Exemplos de sistemas

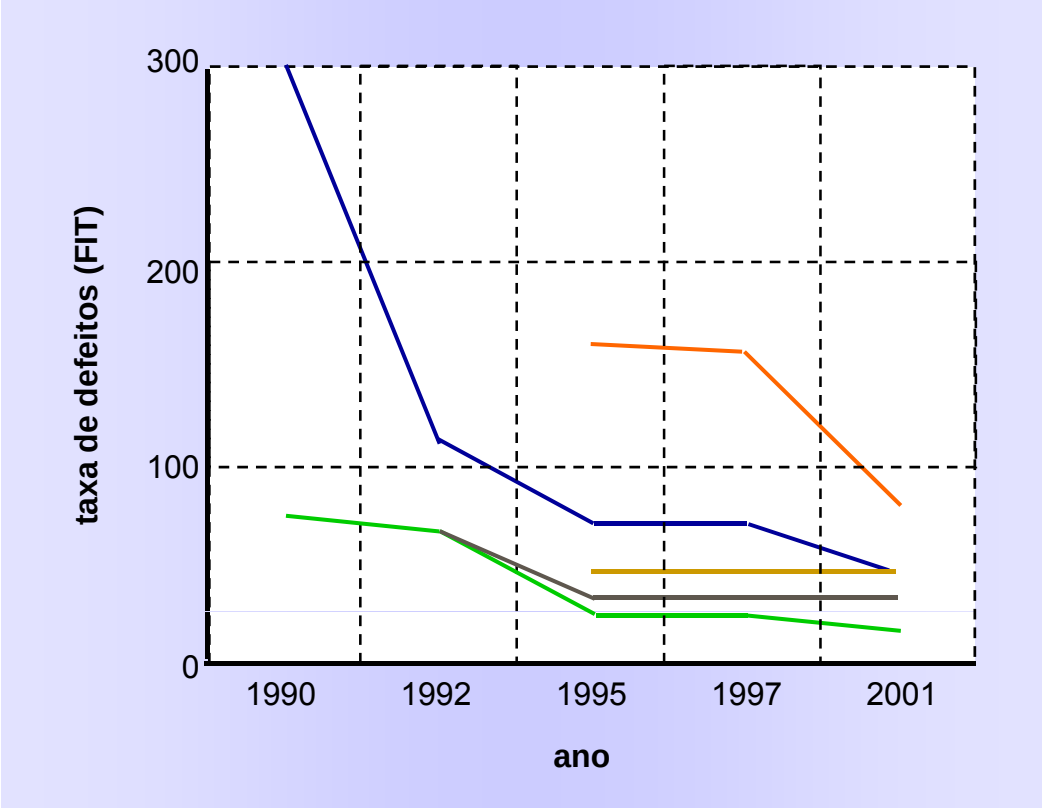


COTS

- componentes comerciais
  - produzidos em larga escala
  - baixo custo
  - facilidade de obtenção
  - baixa taxa de defeitos (para falhas permanentes)
    - FIT baixo e com tendência a diminuir
      - (FIT = failures per 10^9 hours)
  - muito suscetíveis a falhas transientes

desafio - construir sistemas tolerantes a falhas com componentes não confiáveis

Falhas permanentes em COTS



taxa de defeitos (em FIT) devido a falhas permanentes em dispositivos CMOS (COTS) na década de 90 até 2001

FIT = failures per 10^9 hours

Cristian Constantinescu, TRENDS AND CHALLENGES IN VLSI CIRCUIT RELIABILITY. IEEE Micro, 2003

Fontes de defeitos em COTS

- ✓ mas
  - ✓ falhas permanentes não são a principal fonte de preocupação

FIT = failures per 10^9 hours
FIT é medida de taxa de defeito (devido a falhas permanentes)

- ✓ problema maior: falhas transientes
  - ✓ alta susceptibilidade a interferências ambientais

redução no tamanho dos componentes e no nível de potência aumenta a susceptibilidade a interferências por radiação e outros fatores que causam falhas transientes

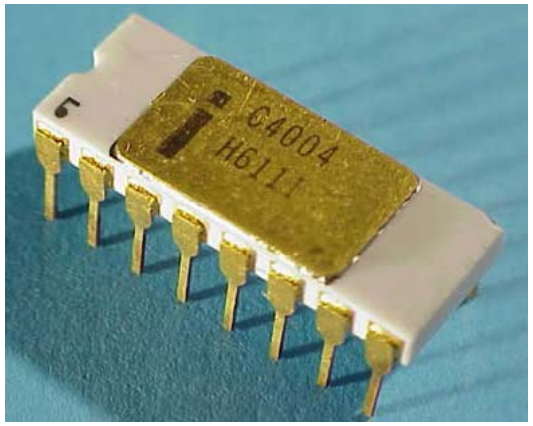
Estudo google sobre memória

- 2 anos
  - servidores com memória protegida por ECC
    - ECC comum (SECDED) e Chipkill
    - chipkill de 4 a 10 vezes mais eficiente que SECDED
  - erros corrigidos
    - mais do que 8% dos chips e 1/3 das máquinas por ano
    - FIT: 25000 a 70000 por Mbit
  - erros não corrigidos
    - 0,22% dos chips e 1,3% das máquinas
  - fortes evidências que erros *hard* são mais comuns que *soft*

Google Inc.: Schroeder, B.; Pinheiro, E.; and Weber, W. "**DRAM Errors in the Wild: A Large-Scale Field Study**." SIGMETRICS/Performance "09, Seattle, WA, June 15-19, 2009.

Desgaste em COTS

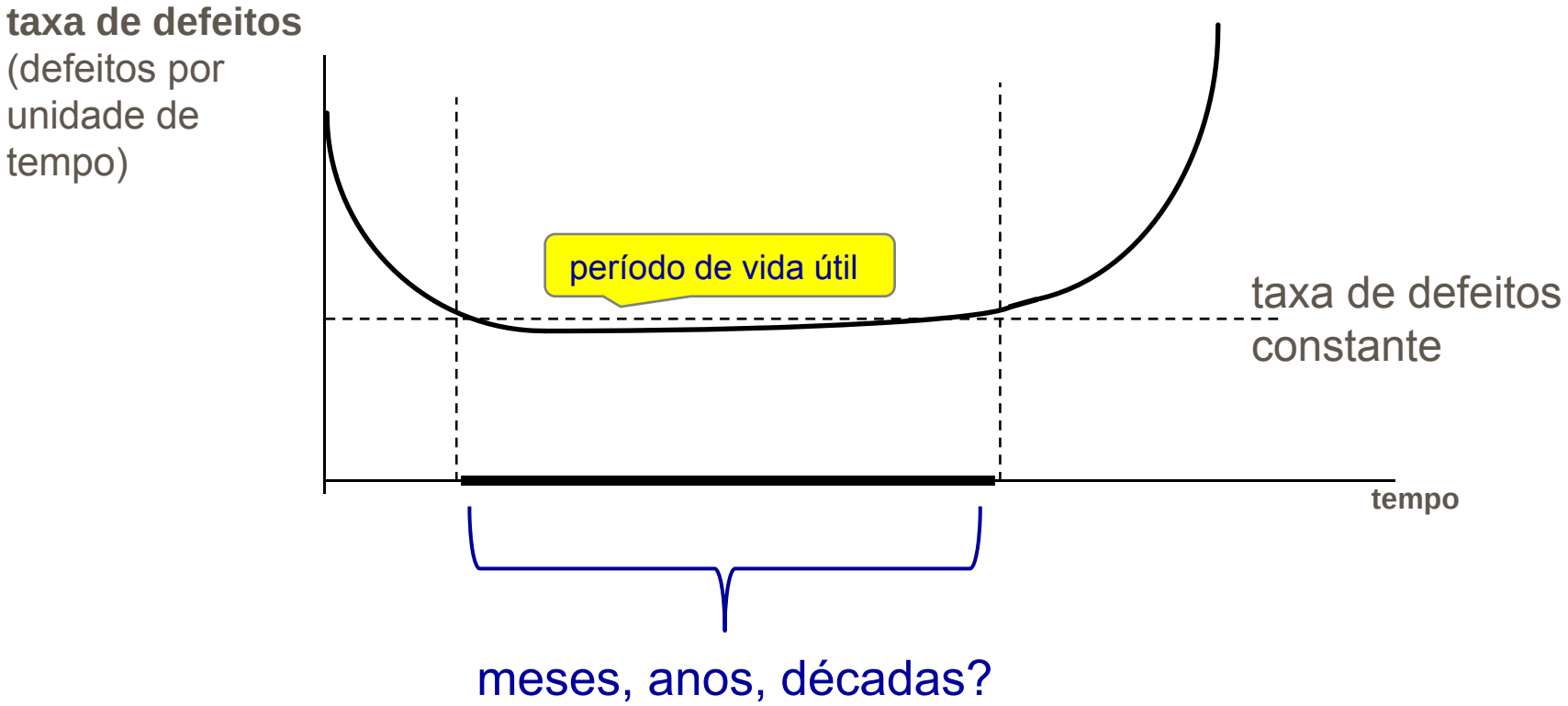
- incerteza sobre desgaste
  - medida de desgaste
    - importante para sistemas fechados em aplicações de vida longa



  - desgaste pode começar após alguns anos
    - ainda sem medidas de envelhecimento
  - taxa de defeitos pode aumentar no tempo
    - FIT baixo e constante pode ser uma medida com validade por curto espaço de tempo

Incerteza quanto a desgaste

qual o tempo de vida útil para o qual vale o FIT baixo e constante?



TF em microprocessadores comerciais

- solução externa:
  - duplicação ou replicação de chips
  - hardware adicional (votadores e comparadores)
- solução interna ao chip:
  - TF suprido pelo próprio microprocessador

basta diminuir o FIT dos microprocessadores e demais chips para garantir um sistema tolerante a falhas?

TF em COTS: exemplos

- Intel
  - desde o 486 na família x86
  - 432 na década de 70

432 era excelente conceitualmente, mas foi um fiasco comercial (alguns afirmam que jamais foi produzido)

sem o emprego de TF, os chips atuais não funcionariam, o MTTF seria muito pequeno

Micros Intel

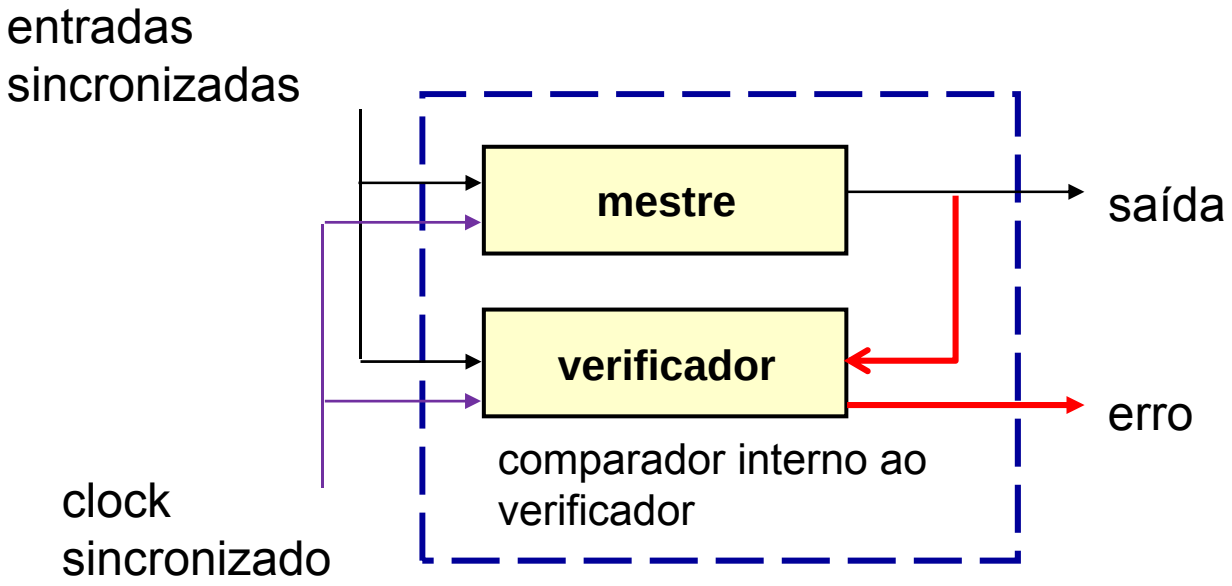
- microprocessadores Intel
  - 100 FITs (FIT = failures per 10^9 hours)
  - MTTF potencial aprox. 1100 anos
- grande MTTF não indica ausência de problemas
  - suscetibilidade a falhas transientes
  - incerteza sobre desgaste
  - numerosas falhas de projeto (errata)
    - Intel P6 (início de 1999): 45 a 101 falhas de projeto
    - novas erratas: taxa de uma por mês

comportamento sob falhas transientes e erratas indicam a necessidade de tolerância a falhas externa ao chip (Avizienis)

Micros Intel - FRC

FRC - Functional Redundancy Checking

mestre e verificador devem estar sincronizados clock-by-clock (lockstep)



Pentiums

paridade: dados, caches, TLB e memória de microcódigo

verificação de exceções: MCA

mestre / verificador FRC

Pentium

Pentium Pro

ECC para dados

2 bits de paridade para linhas de endereço associado a técnicas de retry (re-tentativa)

paridade para sinais de controle

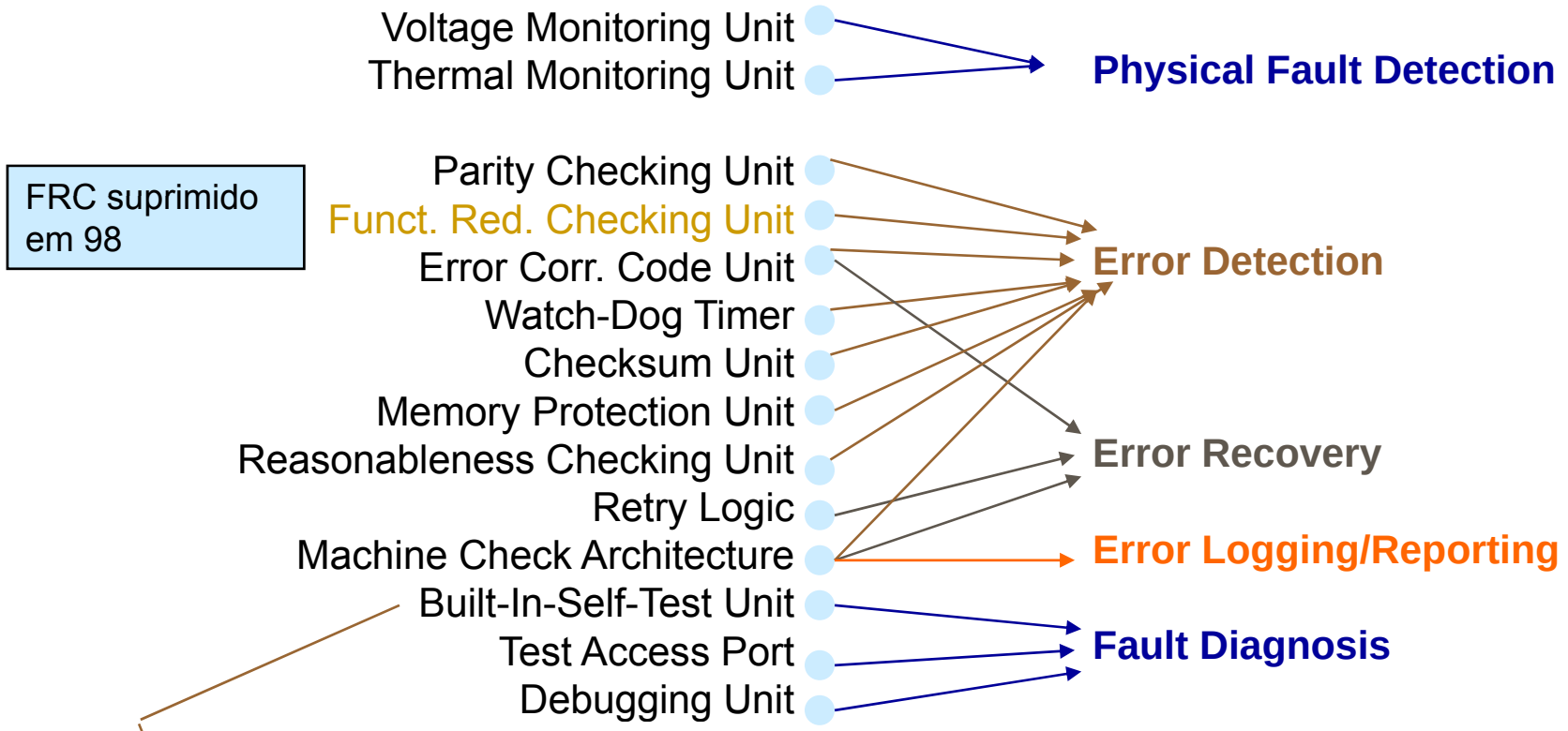
detecção de erros por paridade para caches e barramentos

barramento de dados e memória com ECC

Intel P6

AVIZIENIS, A. **Fault Tolerance Infrastructure for Dependable Computing with High-Performance COTS Components**. DSN, IEEE 2000

Pentium II


Voltage Monitoring Unit
Thermal Monitoring Unit
Physical Fault Detection
FRC suprimido em 98
Parity Checking Unit
Funct. Red. Checking Unit
Error Corr. Code Unit
Watch-Dog Timer
Checksum Unit
Memory Protection Unit
Reasonableness Checking Unit
Retry Logic
Machine Check Architecture
Error Detection
Error Recovery
Error Logging/Reporting
Built-In-Self-Test Unit
Test Access Port
Debugging Unit
Fault Diagnosis
para teste dos elementos de memória interna (microcódigo, caches, TLB)


pentium II
intel

Itanium 2

- ✓ recuperação de erros de barramento de dados
- ✓ cache ECC (já existente no P6)
- ✓ correção de erro simples de memória
- ✓ re-tentativa na detecção de erro duplo de memória
- ✓ suporte a memória espelhada (spare)
- ✓ verificação de erros soft (transientes) na lógica interna: bit de paridade
- ✓ suporte a *lockstep*
- ✓ contenção de dados corrompidos

memória com defeito é substituída por memória estepe

marca a porção de memória com dados corrompidos e limita o uso dos dados a apenas um programa;
elimina os dados quando o programa termina ou sobrescreve a porção

Reliability, Availability, and Serviceability for the Always-on Enterprise. White paper, Intel, 2005

MCA

- machine check architecture
  - registradores dedicados para log de erros
    - facilita diagnóstico
  - capacidade de manipulação de erros

P6

MCA opcional - pode ser desligado por software

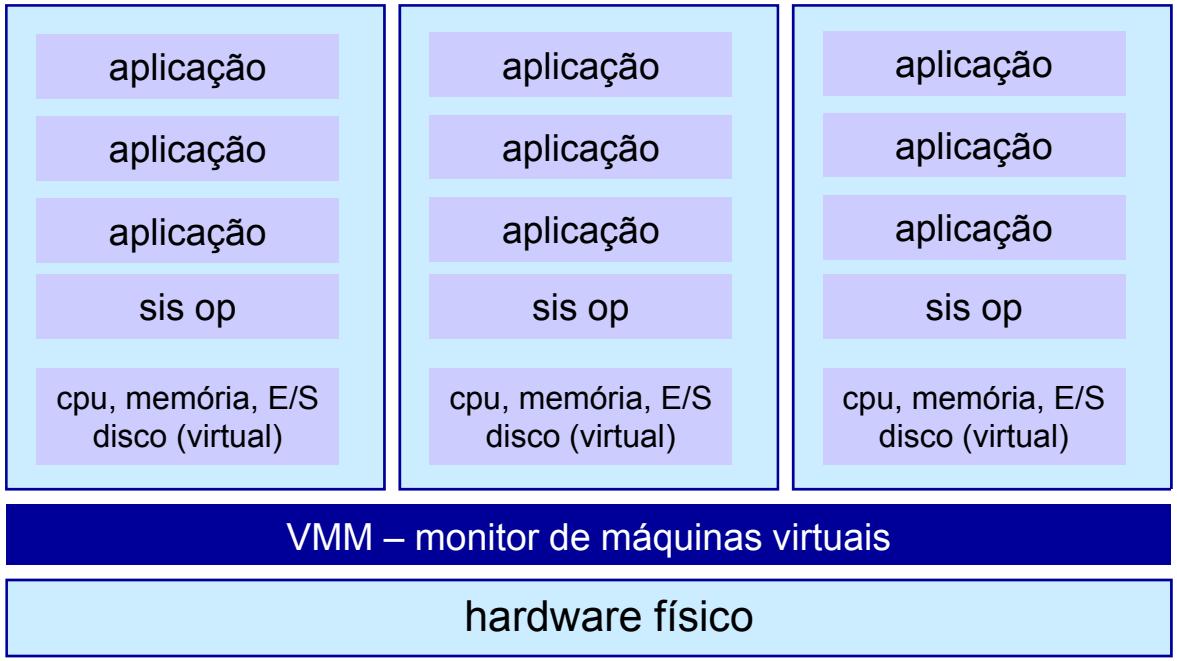
- Advanced MCA
  - segue padrões
    - padrões facilitam a interface com o SO e firmware
  - permite ao SO e ao firmware recuperarem erros complexos
  - pode *resetar* o sistema automaticamente em resposta a erros fatais

Itanium 2

Virtualização Intel

Enhanced Virtualization on Intel® Architecture based Servers - White paper, Intel 2004

VM (máquina virtual):
isolamento de falhas por hardware
possibilidade de implementar *failover* na mesma máquina

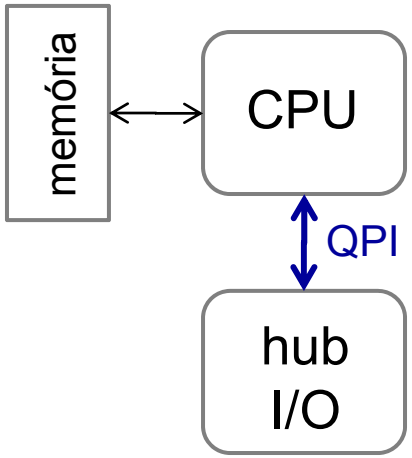


novo anel de privilégio (maior que kernel) para a VMM

VMM – software que opera como árbitro no acesso aos recursos físicos e hospeda as máquinas virtuais

Intel i7 e Xeon

- ✓ Intel® QuickPath Interconnect (QPI)
  - ✓ detecção de erro com CRC
  - ✓ correção de erro usando *Link level retry*
  - ✓ Intel® Interconnect Built In Self Test



- ✓ MCA Recovery
  - ✓ permite recuperação do sistema
  - ✓ sinaliza o erro para o SO ou VMM que podem então recuperar o erro (se for possível) sem derrubar todo o sistema

herança do Itanium

Intel White Paper: Intel® Xeon® Processor E7 Family: Reliability, Availability, and Serviceability - Advanced data integrity and resiliency support for mission-critical deployments. 2011

Intel Xeon processor E7

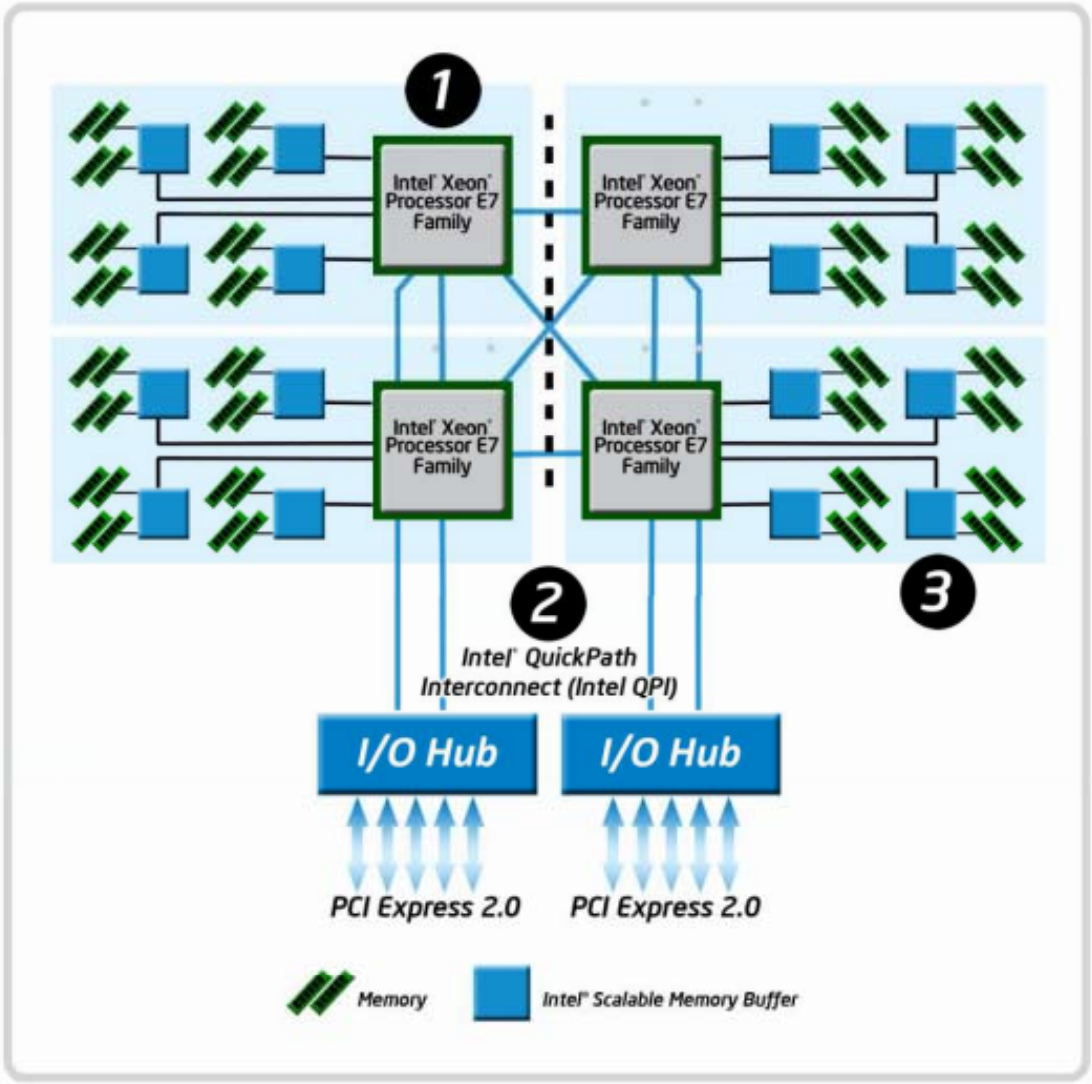
1 PROCESSOR/SYSTEM

- Corrupt Data Containment Mode
- Electronically Isolated Partitioning
- Processor Sparing and Migration*
- Core (Socket) Disable for Fault Resilient Boot
- Machine Check Architecture Recovery (MCA Recovery)*
- CPU Hot Add*
- PCIe Express Hot Plug
- Corrected Machine Check Interrupt (CMCI) for Preventive Failure Analysis*

*Requires operating system support.

2 INTEL® QPI

- Intel QPI Protocol Protection via CRC
- QPI Viral Mode
- Intel QPI Self-healing
- QPI Clock Failover
- QPI Packet Retry



3 MEMORY

- ECC
- Memory Address Parity Protection
- Memory Demand and Patrol Scrub
- Memory Thermal Throttling
- Enhanced DRAM Single Device Data Correction (SDDC)
- Enhanced DRAM Double Device Data Correction (DDDC+1)
- Fine Grained Memory Mirroring
- Memory Sparing
- Memory Migration
- Intel Scalable Memory Interconnect (SMI) Lane Failover
- Intel SMI Clock Failover
- Intel SMI Packet Retry
- Failed DIMM Identification
- Memory Hot Add*

*Requires operating system support.

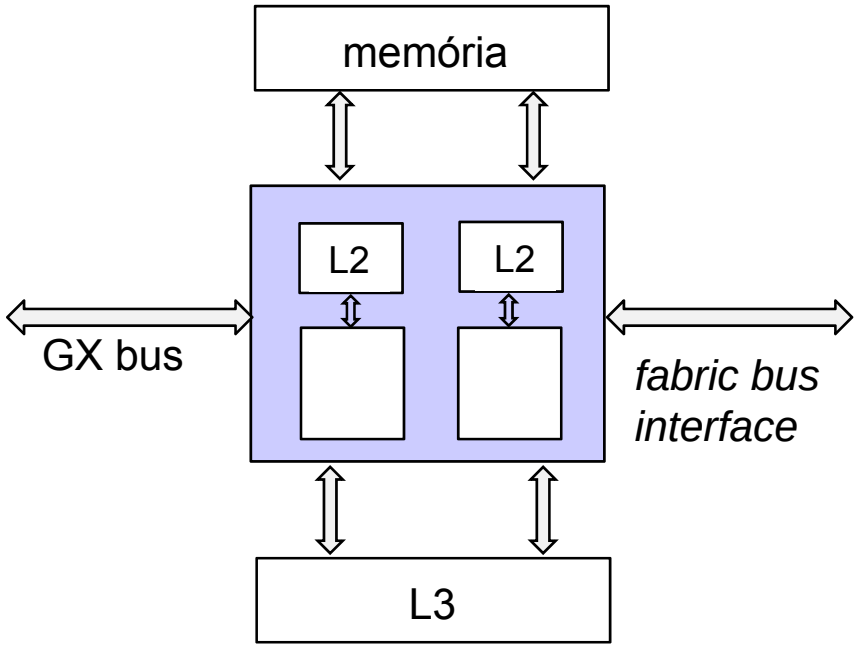
Micros recentes

- ✓ IBM Power5 / Power6/ Power7

- ✓ vários outros
  - ✓ ARM Cortex R Series
  - ✓ SPARC64 V
  - ✓ SUN Niagara II
  - ✓ AMD Opteron
  - ✓ Texas Hercules

IBM Power

- ECC — comum ao Power5 e Power6
  - sinais internos ao chip
  - todos pinos entre o chip e:
    - L3 cache
    - memória
    - GX bus que conecta o chip ao I/O hub
    - *fabric bus interface* para comunicação entre chips e entre nodos
  - L2 e L3 caches
  - memória



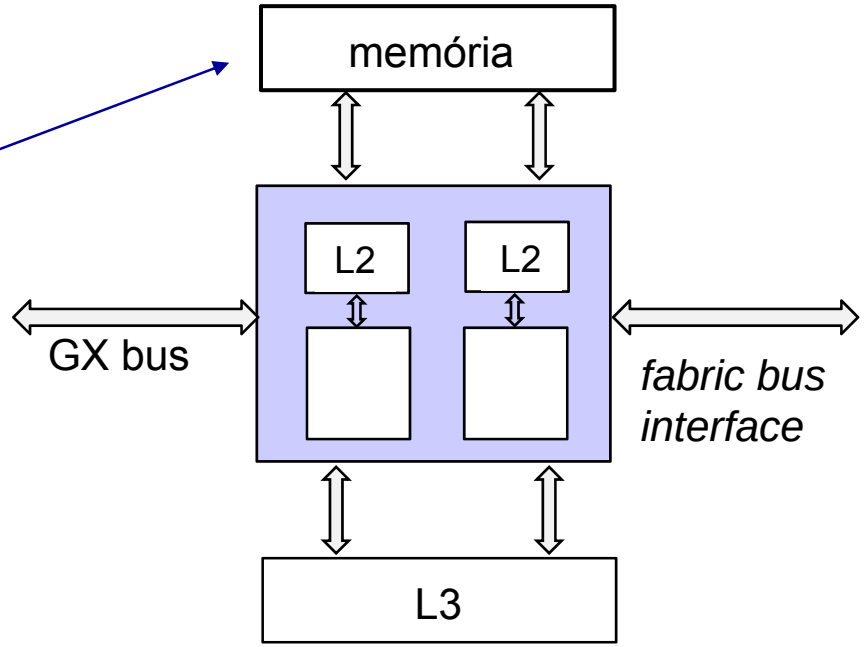
POWER6™ Built on Power™

as L1 são internas aos cores e protegidas por **paridade**

IBM Power

comum ao Power 5 e 6

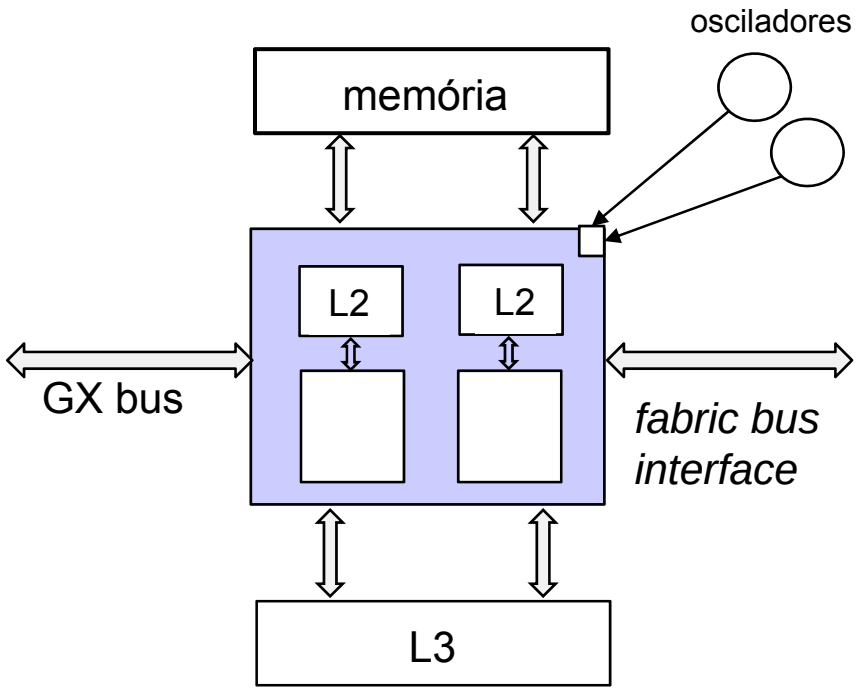
- ChipKill
- *scrubbing* (limpeza) assistida por hardware
- redundância dinâmica e *bit steering* (troca por estepe)
- tratamento para *special uncorrectable error* (SUE)



IBM Power6

novas técnicas no **Power6**:

- reparo dinâmico no barramento de memória,
- failover dinâmico de oscilador (clock)
- recuperação de cache
- recuperação para outro processador
- partition isolation for core checkstops
- *instruction retry* para erros detectados nos cores



IBM Power6 - RU

- unidade de recuperação - RU
  - faz **checkpoint** do estado do sistema no final da execução de um grupo de instruções
  - array de checkpoints – protegido por ECC
    - cuidados especiais são tomados para que o checkpoint não registre erros
- recuperação é realizada para todo o core
  - não apenas para uma thread
  - um pequeno período de tempo após a recuperação, o core fica sem operação em pipeline (slow mode)

Jude A. Rivers and Prabhakar Kudva. Reliability Challenges and System Performance at the Architecture Level. IEEE Design & Test of Computers. 2009. p. 62-72

FAULT-TOLERANT DESIGN OF THE IBM POWER6 MICROPROCESSOR" , K Reick, PN Sanda, S Swaney, JW Kellington, Michael Mack, Michael Floyd e D. Henderson ,publicado na IEEE Micro, 2008, pg. 30 a 38

IBM Power6 - IRR

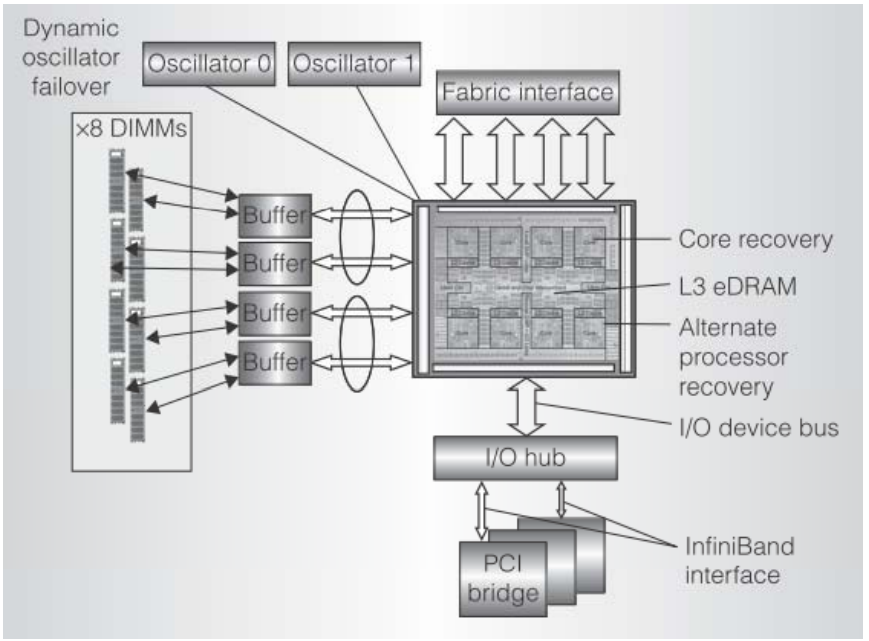
- IRR
  - instruction retry recovery
  - o RU (recovery unit) dispara o IRR
  - se foi falha transitória, a recuperação é bem sucedida
  - se foi permanente, é escalado um checkstop
    - no checkstop vai ser feita uma recuperação de alto nível para um core alternativo

Jude A. Rivers and Prabhakar Kudva. Reliability Challenges and System Performance at the Architecture Level. IEEE Design & Test of Computers. 2009. p. 62-72

FAULT-TOLERANT DESIGN OF THE IBM POWER6 MICROPROCESSOR" , K Reick, PN Sanda, S Swaney, JW Kellington, Michael Mack, Michael Floyd e D. Henderson ,publicado na IEEE Micro, 2008, pg. 30 a 38

Power 7

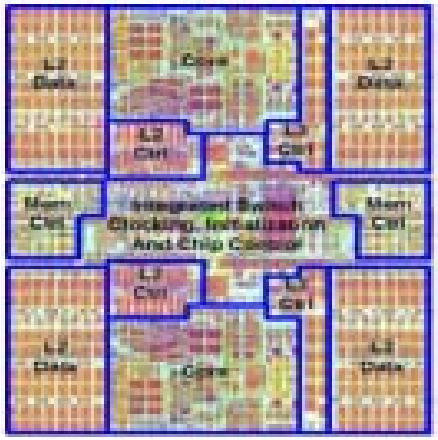
- ✓ todas as características RAS do Power 6
- ✓ adicionais
  - ✓ algoritmo ECC 64 bits para a memória
    - ✓ permite a correção de 8 bits (ou seja um chip de memória)
  - ✓ estepe para os chips de buffer de memória
  - ✓ espelhamento de memória seletivo



Ron Kalla, B. Sinharoy, W. J. Starke, M. Floyd. POWER7: IBM'S NEXT-GENERATION SERVER PROCESSOR. IEEE Micro. 2010.

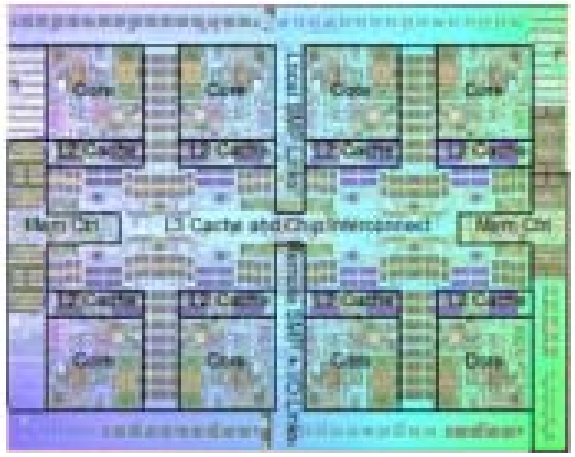
POWER7 System RAS - Key Aspects of Power Systems Reliability, Availability, and Serviceability
October 3, 2012 , IBM Systems and Technology Group , Daniel Henderson, Jim Mitchell, and George Ahrens

Power6 e Power7



POWER6 (2007)

- 65 nm technology – 341 mm^2
- 0.79B transistors
- 2 Cores
  - 2 SMT threads/core
- 9 execution units/core
  - 2 integer and 2 binary floating-point units
  - 1 vector and 1 decimal floating-point unit
  - 2 load/store, 1 branch
- Integrated L2 cache
- L3 directory & controller (off chip L3 cache)
- 2 memory controllers



POWER7 (2010)

- 45nm technology – 567 mm^2
- 1.2B transistors
- 8 Cores
  - 4 SMT threads/core
- 12 execution units/core
  - 2 integer and 4 binary floating-point units
  - 1 vector and 1 decimal floating-point unit
  - 2 load/store, 1 branch, 1 condition register
- Integrated L2 cache
- Integrated L3 cache
- 2 memory controllers

Bibliografia

- capítulos de livros
  - SIEWIOREK, D. Architecture of fault-tolerante computers, cap 2. **Fault-Tolerant System Design**. Prentice Hall, New Jersey, 1996
- livros
  - SURI, N.; WALTER, C.J.; HUGUE, M.M. **Advances in ultra-dependable distributed systems**. IEEE Computer Society Press. Los Alamitos. 1995.

Bibliografia

- artigos
  - AVIZIENIS, A**. Fault Tolerance Infrastructure for Dependable Computing with High-Performance COTS Components**. DSN, IEEE 2000
  - Cristian Constantinescu, **Trends and Challenges in VLSI Circuit Reliability. IEEE Micro, 2003**
  - R. Iyer et al. **Recent Advances and New Avenues in Hardware-Level Reliability Support**, IEEE Micro, vol. 25, pp. 18-29, 2005.
  - K Reick, PN Sanda, S Swaney, JW Kellington, Michael Mack, Michael Floyd e D. Henderson. **Fault-tolerant Design of The IBM Power6 Microprocessor**. IEEE Micro, 2008, pg. 30 a 38
  - Daniel Henderson, Jim Mitchell, and George Ahrens **POWER7 System RAS Key Aspects of Power Systems Reliability, Availability, and Serviceability .** October 3, 2012. IBM Systems and Technology Group

Bibliografia

- artigos
  - Ron Kalla, Balaram Sinharoy, William J. Starke, Michael Floyd**. POWER7: IBM'S NEXT-GENERATION SERVER PROCESSOR**. IEEE Micro. 2010. March/ april. 7-15
  - Google Inc.; Schroeder, Bianca; Pinheiro, Eduardo; and Weber, Wolf-Dietrich. "**DRAM Errors in the Wild: A Large-Scale Field Study**." SIGMETRICS/Performance ‟09, Seattle, WA, June 15-19, 2009.
  - Intel White Paper: Intel® Xeon® Processor E7 Family: Reliability, Availability, and Serviceability - Advanced data integrity and resiliency support for mission-critical deployments. 2011