BLUE WATERS

SUSTAINED PETASCALE COMPUTING

July 25, 2011

Blue Waters, PetaFlops e o Futuro

Celso L. Mendes

Univ. Illinois – Dep. Computer Science

INPE – Lab. Assoc. Computação e Matemática Aplicada

NCSA

IBM

GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

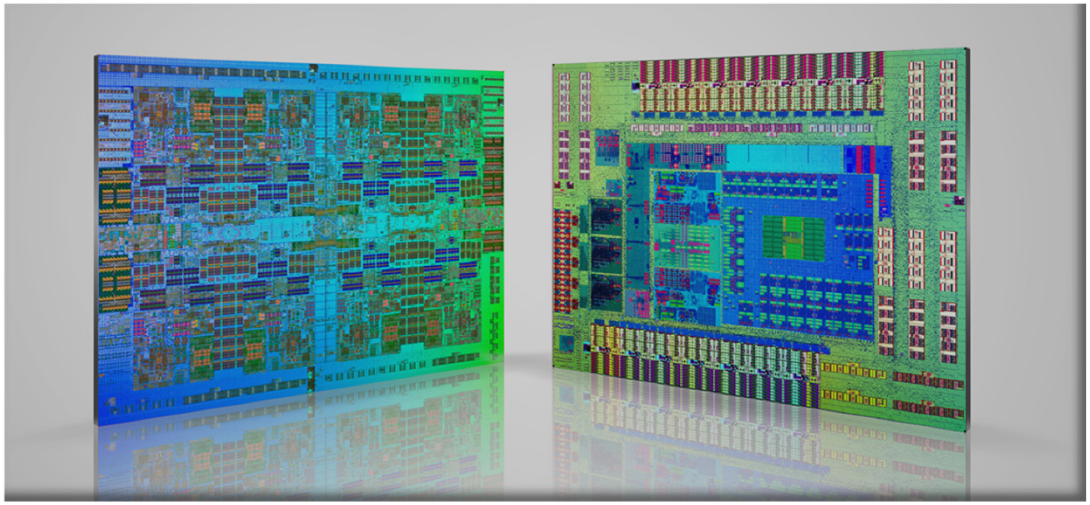
Tópicos:

1. Sistema Blue Waters
   a) Hardware: CPU, Interconexão
   b) Software: Suporte, Applicações
2. Sistemas Petascale Atuais
3. Próxima Fronteira: Sistemas Exaflop
4. Outras Atividades Correntes
   a) Na Universidade de Illinois
   b) Em São José dos Campos

Sistema Blue Waters - Origem

- Edital NSF-06-573: Junho-2006 - $200 Milhões
  - Requisito: 1 Petaflop/s *efetivo* em três aplicações
    - Dinâm. molecular, turbulência, cromodin. quantica
- Propostas submetidas: Set.2006 (pré), Fev.2007
- Escolha final do vencedor: Agosto-2007
  - Blue Waters: Univ. Illinois (NCSA) + IBM
  - Sistema baseado em processador Power7
  - Contrapartida de Illinois: prédio, pessoal
  - Instalação prevista: 2011

Coração do Blue Waters: 2 Novos Chips



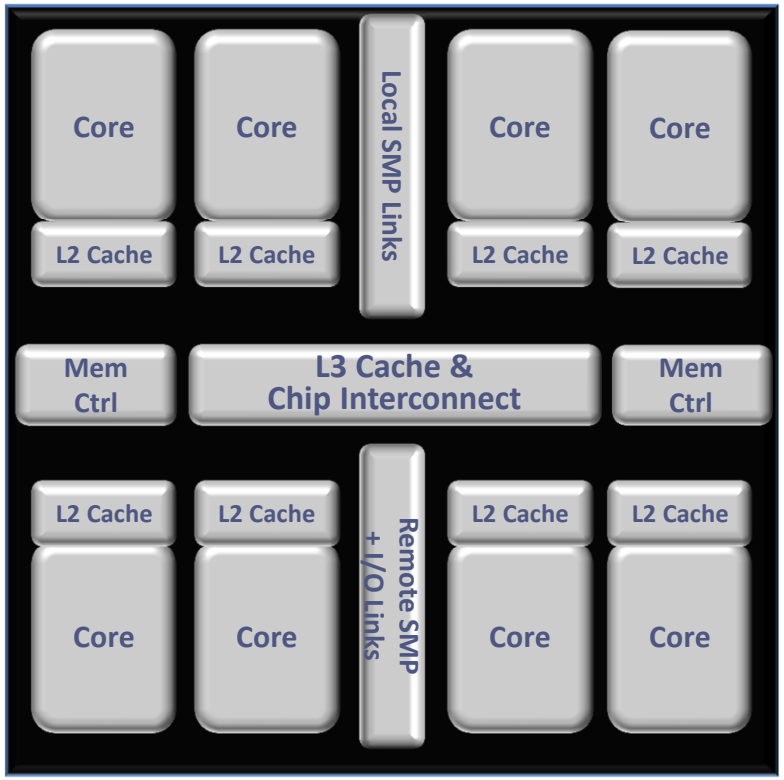
Proc.: IBM Power7

- Até 256 Gflop/s de pico
- Clock: 3.5–4.0 GHz
- Até 8 núcleos, 4 threads/núcleo
- Dois controladores de memória
- Acesso à memória: 128 GB/s

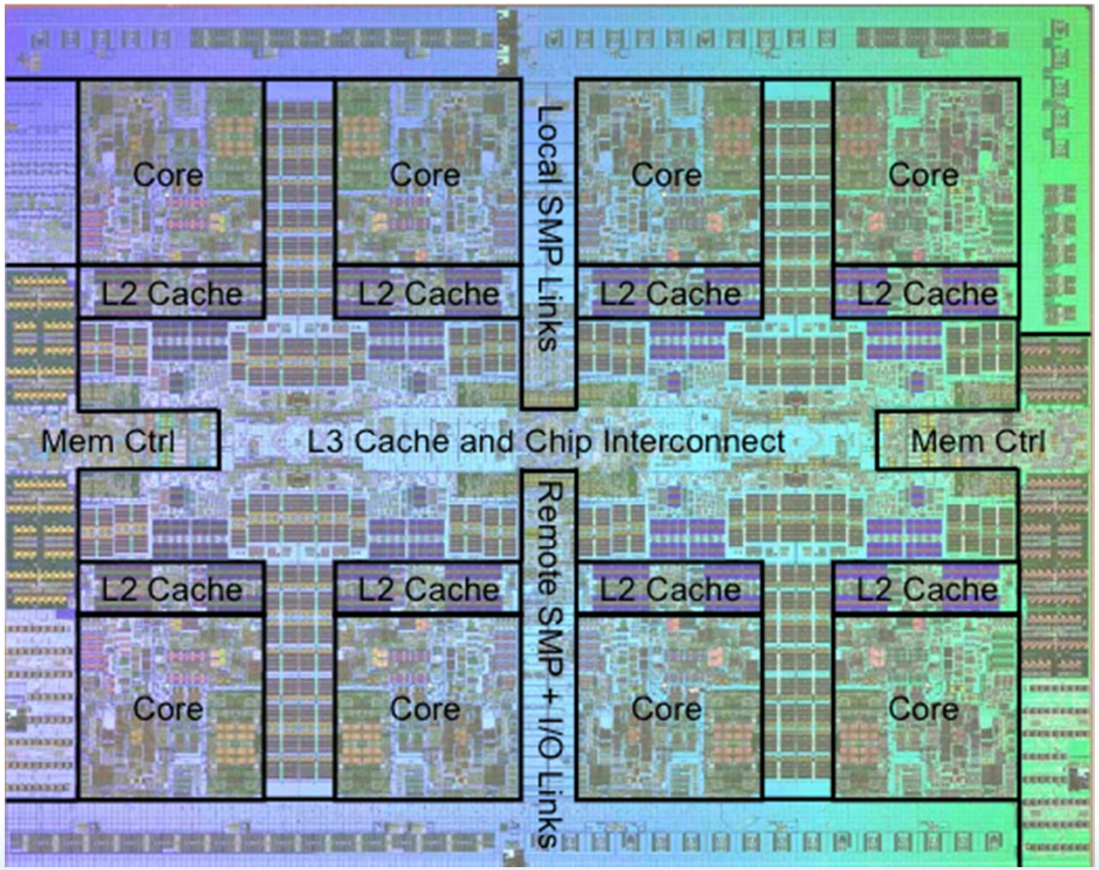
Rede: IBM Hub

- 192 GB/s para as CPUs (até 32)
- 336 GB/s para 7 hubs locais (7 × 24)
- 240 GB/s para 24 hubs semi-locais
- 320 GB/s para hubs distantes
- ~1.12 TB/s largura de banda total!

Chip do Power7: Diagrama de Blocos



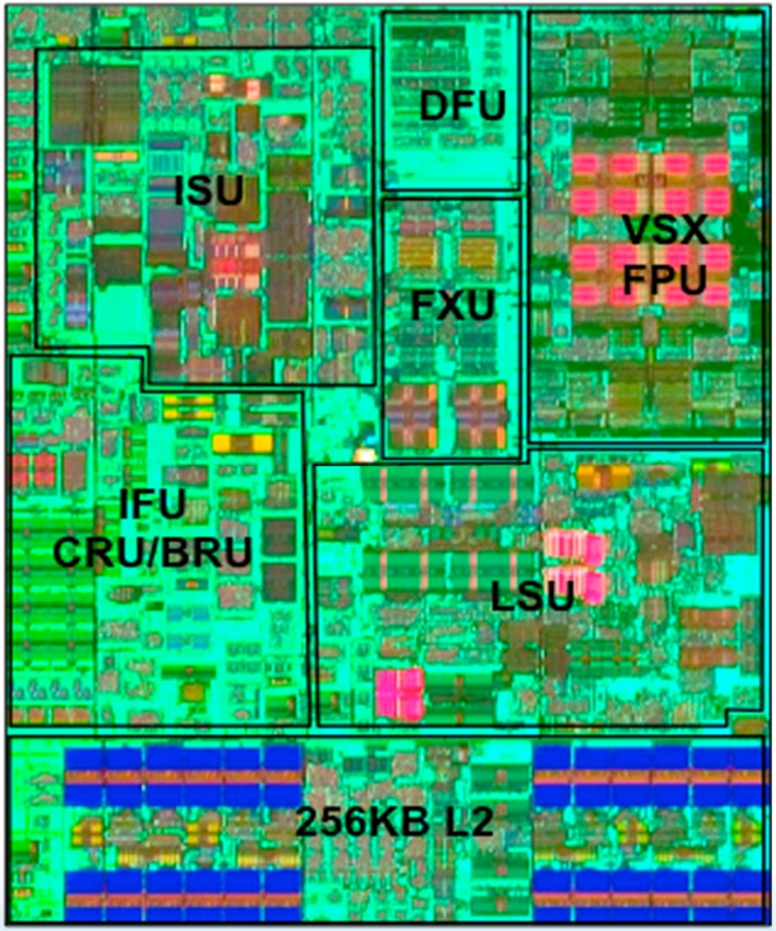
Chip do Power7: 8 Núcleos



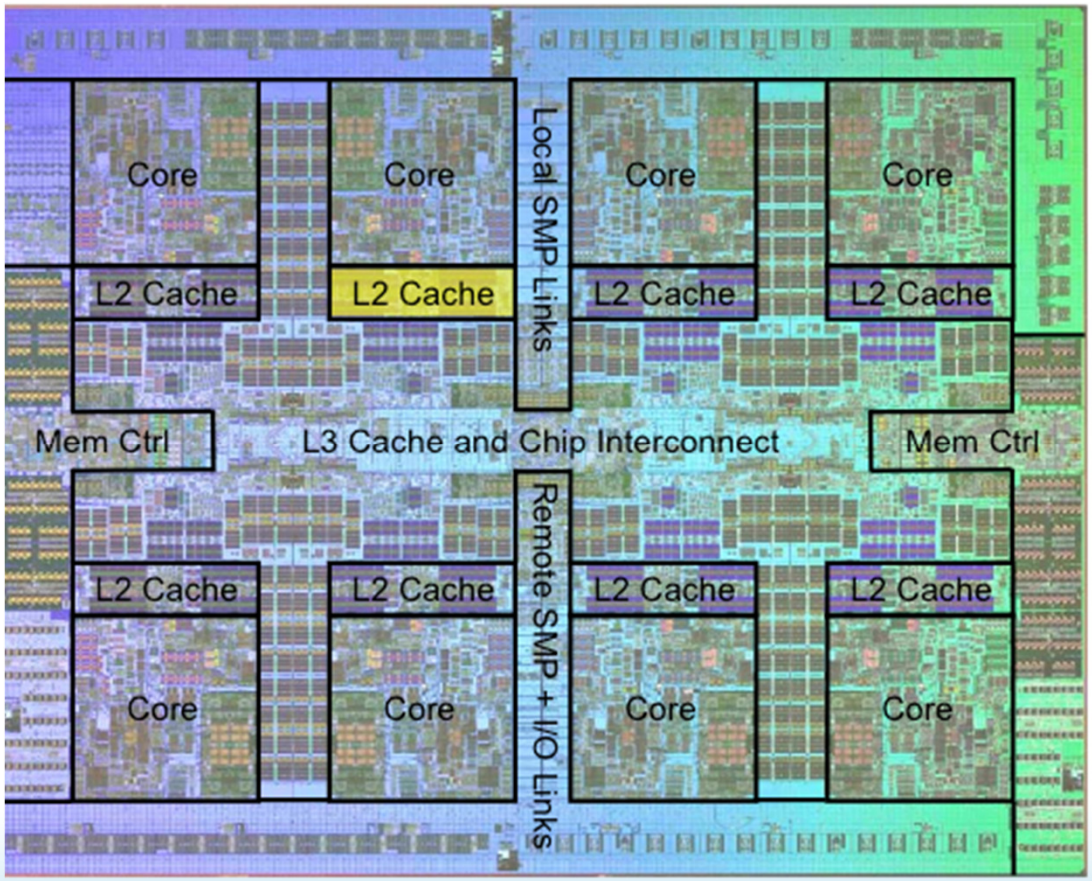
Núcleo de Processamento no Power7

POWER7: Core

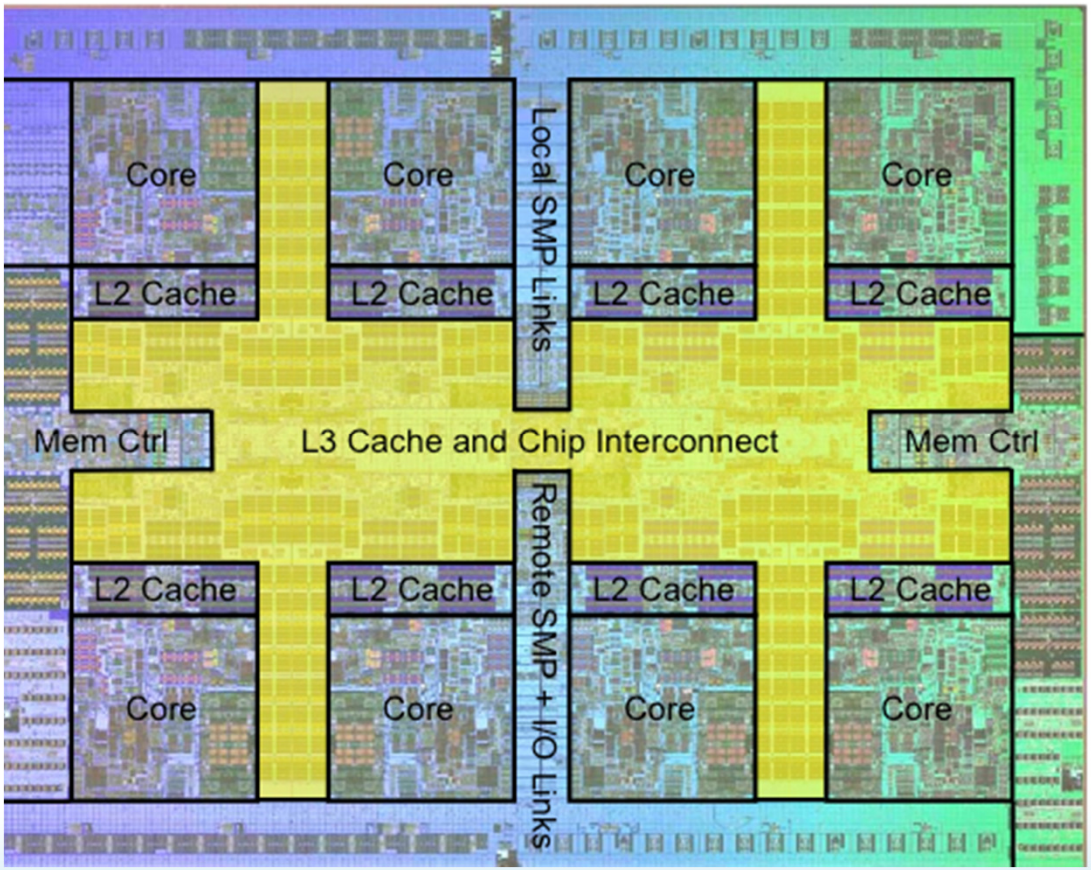
- Execution Units
  - 2 Fixed point units
  - 2 Load store units
  - 4 Double precision floating point
  - 1 Branch
  - 1 Condition register
  - 1 Vector unit
  - 1 Decimal floating point unit
  - 6 wide dispatch
- Recovery Function Distributed
- 1,2,4 Way SMT Support
- Out of Order Execution
- 32KB I-Cache
- 32KB D-Cache
- 256KB L2
  - Tightly coupled to core



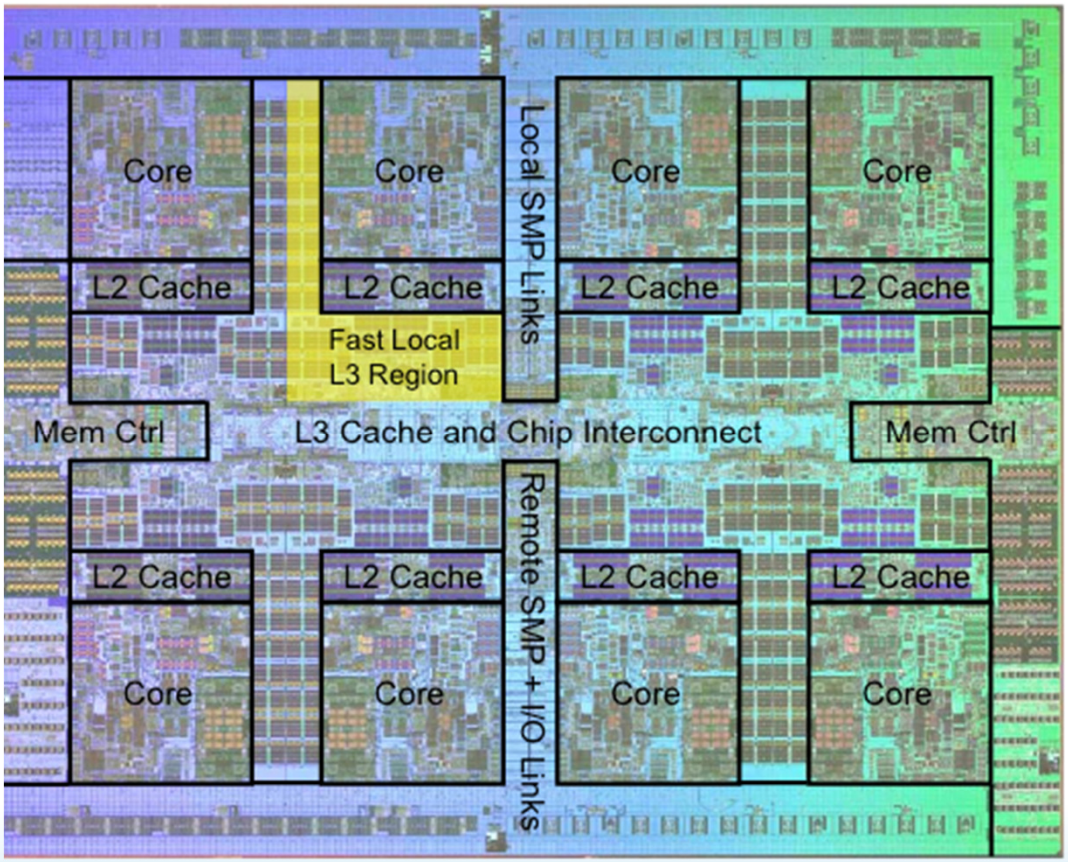
Cache L2: 256KB (D/I)



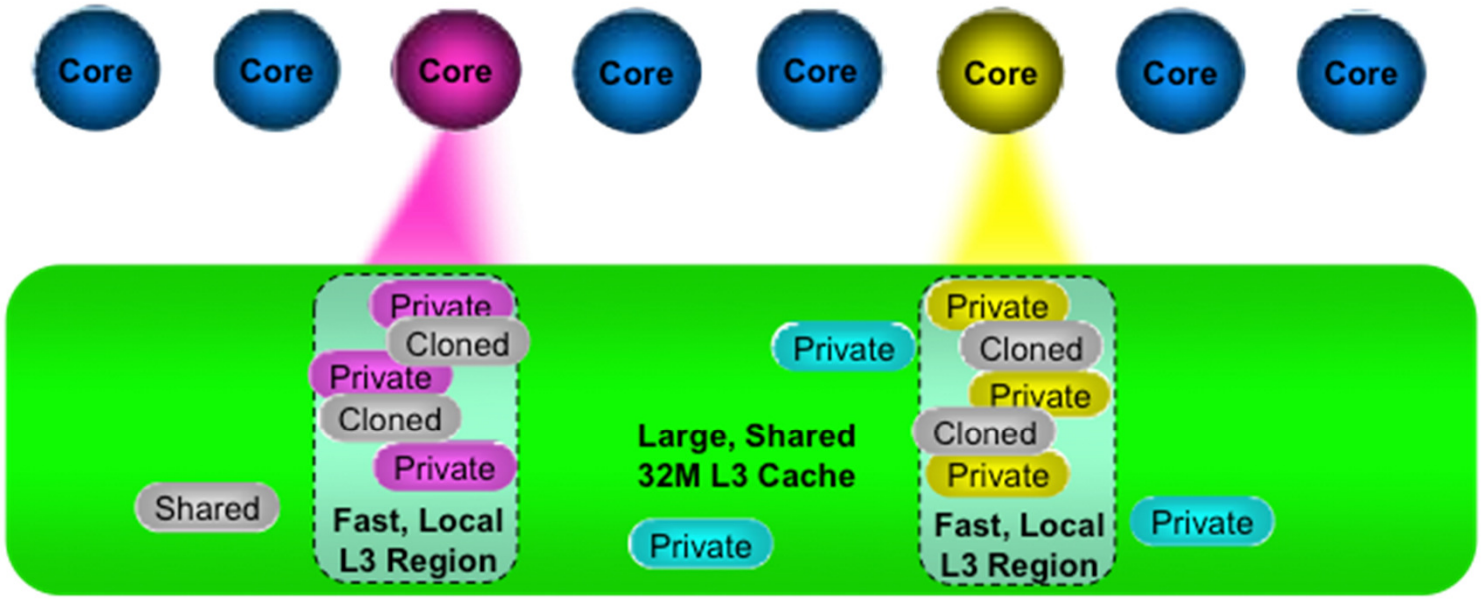
Cache L3: 32 MB Total (8 x 4 MB)



Cache L3: 4 MB "Local" por Núcleo



Cache L3 - Vantagens



- 32 MB de armazenamento interno ao chip Power7
- Velocidade ~3 vezes maior que em acessos a uma memória cache externa
- Cada núcleo pode utilizar toda a L3, se necessário
- Acesso local ~5 vezes mais rápido que acessos à area compartilhada

Desdobramentos do Projeto Power7

Diversos Produtos Comerciais IBM (ex: 770)

- Variável número de núcleos por chip: 4, 6, 8
- Várias velocidades de relógio

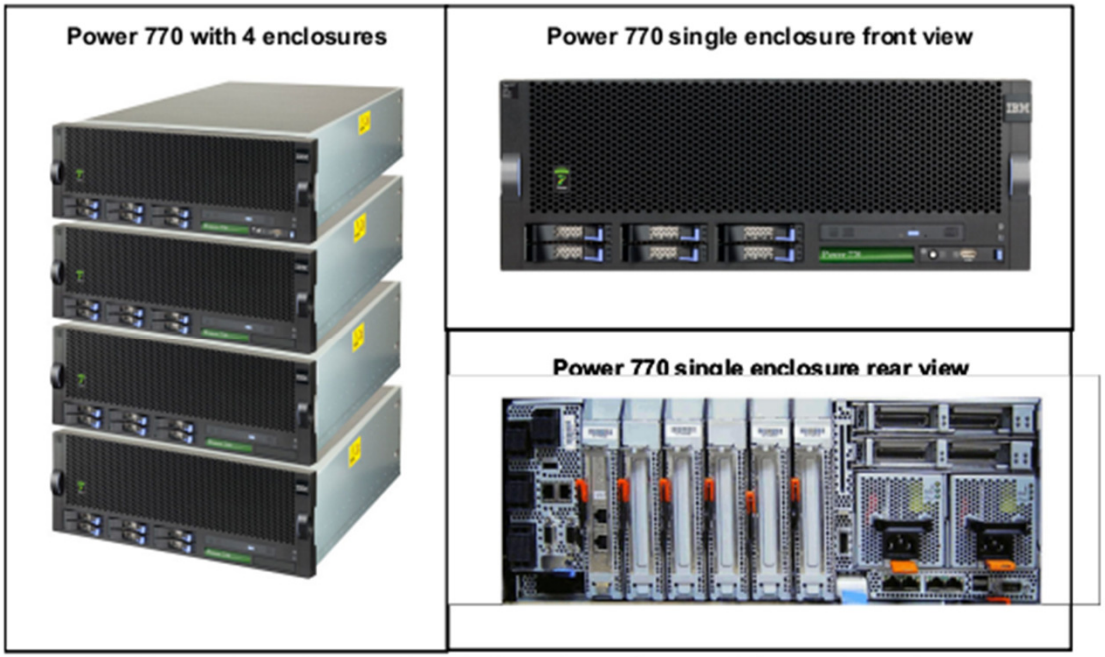
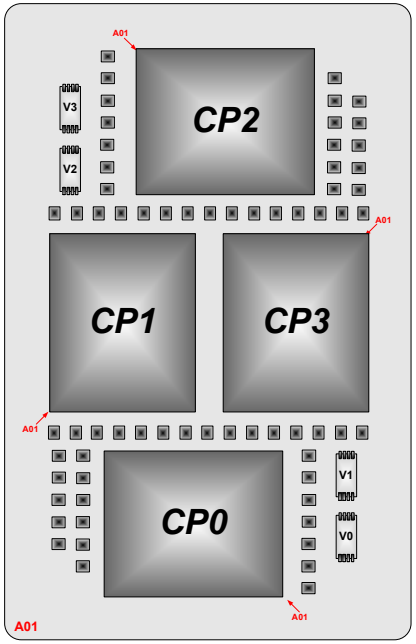
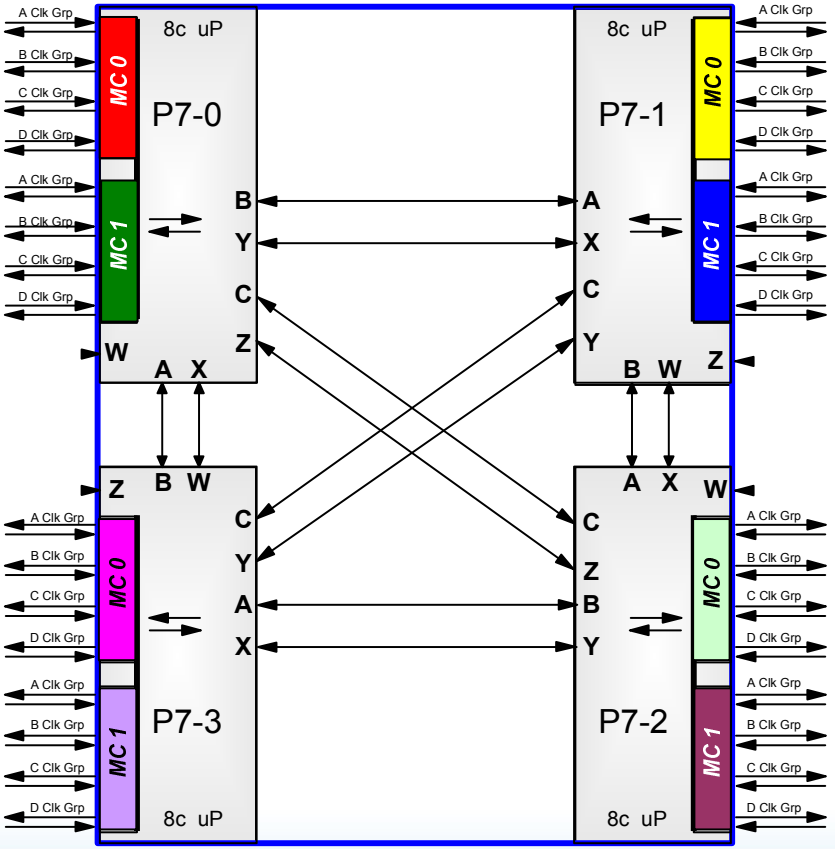


Figure 1-1 Four-enclosure Power 770, a single-enclosure Power 770 front and rear views.

Fonte: IBM

Quad-Chip Module (QCM): 4 Chips Power7

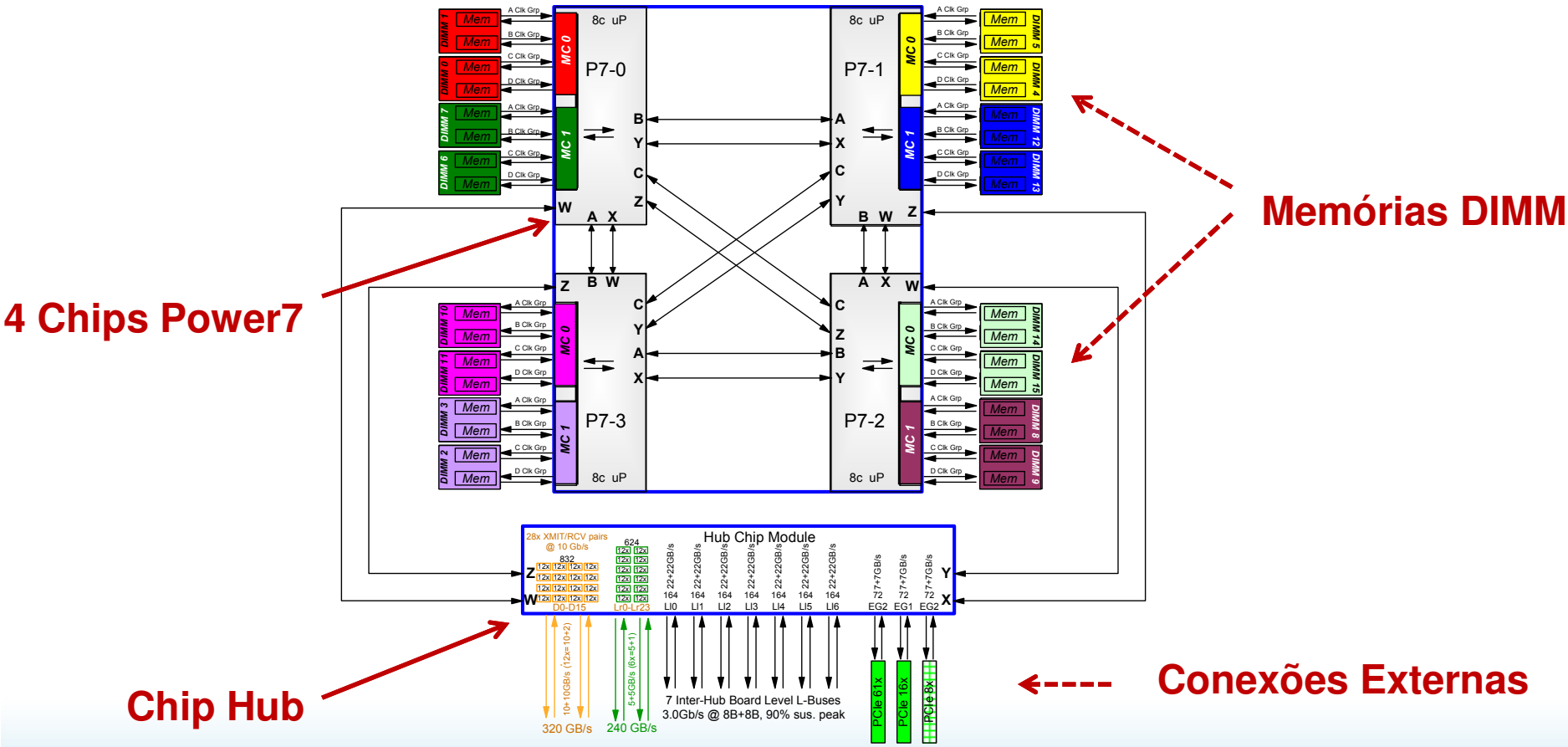




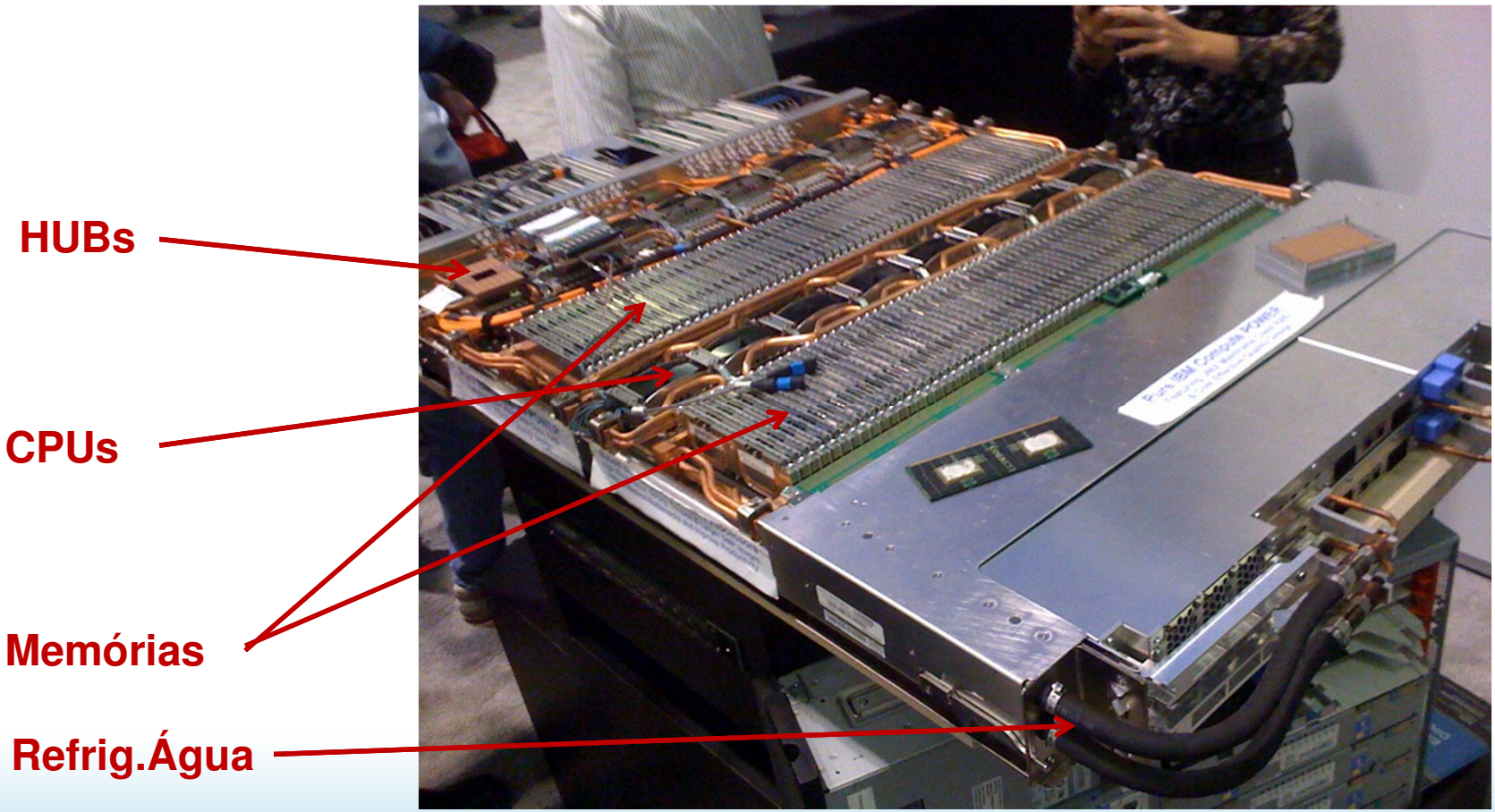
Características do QCM

- 32 núcleos, operando um nó SMP
- Até 4 threads por núcleo (SMT)
- 8 Flops/cycle por núcleo – 4 mult/add
- 3.5 ~ 4.0 GHz
- Pico Teórico: 1 Teraflop/s
- Até 512 GB de memória
- 512 GB/s p/ memória: 0.5 Bytes/Flop

Nó Computacional: QCM + Mem. + Hub

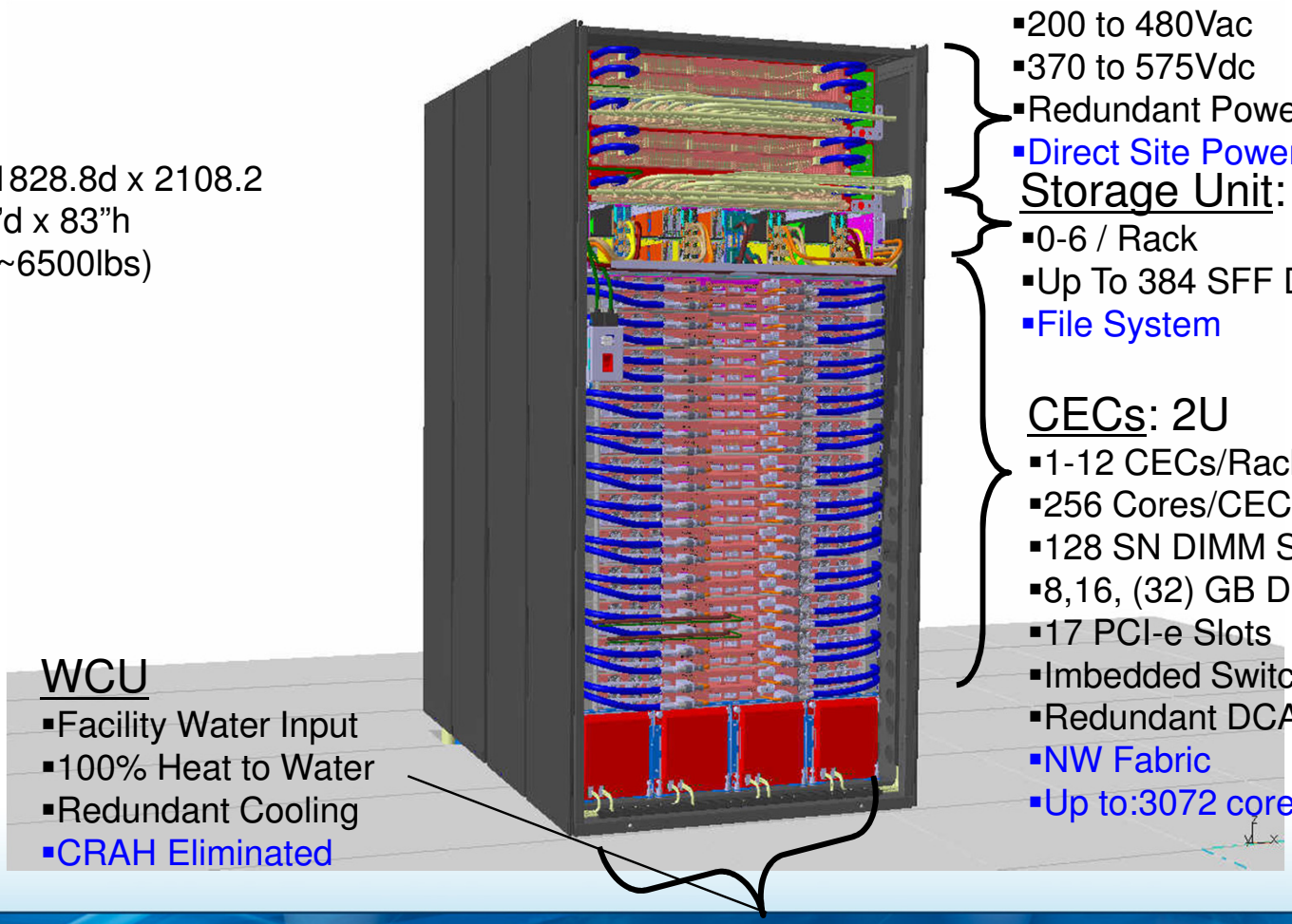


Montagem Física: Gaveta (8 Nós)



Montagem Física: Rack (até 12 Gavetas)

Rack
- 990.6w x 1828.8d x 2108.2
- 39”w x 72”d x 83”h
- ~2948kg (~6500lbs)



BPA
- 200 to 480Vac
- 370 to 575Vdc
- Redundant Power
- Direct Site Power Feed

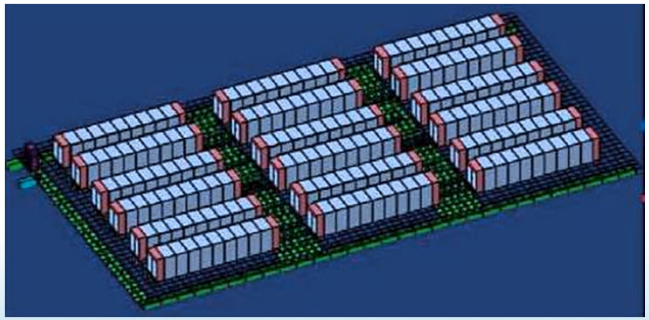
Storage Unit: 4U
- 0-6 / Rack
- Up To 384 SFF DASD / Unit
- File System

CECs: 2U
- 1-12 CECs/Rack
- 256 Cores/CEC
- 128 SN DIMM Slots / CEC
- 8,16, (32) GB DIMMs
- 17 PCI-e Slots
- Imbedded Switch
- Redundant DCA
- NW Fabric
- Up to:3072 cores, 24.6TB

WCU
- Facility Water Input
- 100% Heat to Water
- Redundant Cooling
- CRAH Eliminated

Montagem Física: Blocos Básicos

- Bloco Básico: 3 Racks
  - 2 Racks com 12 gavetas computacionais (CEC)
  - 1 Rack com 8 gavetas computacionais (CEC) e gavetas de armazenamento
  - 8.192 núcleos por bloco básico (262 TFlop/s)
  - 3 blocos básicos por "fila":



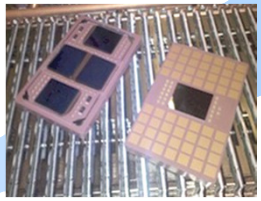
Fonte:
M.J.Ellsworth - IBM

OBS: configuração final a ser ainda determinada

Blue Waters: Modularidade

Chip Power7

8 cores, 32 threads
L1, L2, L3 cache (32 MB)
Up to 256 GF (peak)
128 Gb/s memory bw
45 nm technology



Quad-chip Module

4 Power7 chips
1 TF (*peak*)
128 GB memory
512 GB/s memory bw

Hub Chip

1.128 TB/s comm bw

IH Server Node

8 QCM's (256 cores)
8 TF (*peak*)
1 TB memory
4 TB/s memory bw
8 Hub chips
9 TB/s comm bw
Power supplies
PCIe slots

Fully water cooled

Blue Waters
3-Rack Building Block

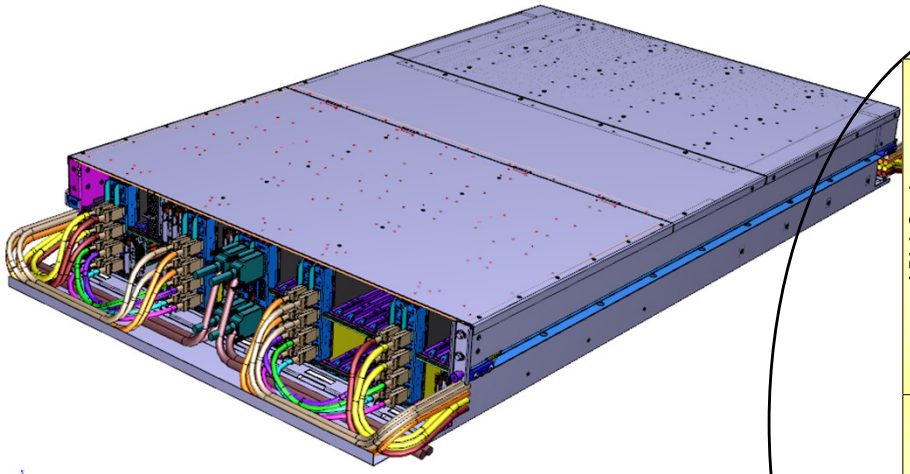
32 IH server nodes
256 TF (peak)
32 TB memory
128 TB/s memory bw
4 Storage systems (>500 TB)
10 Tape drive connections

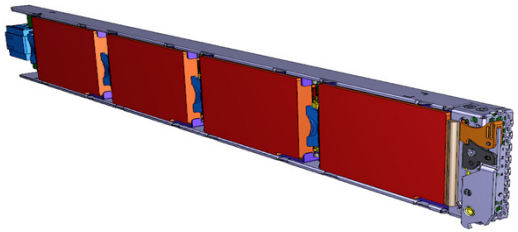
Blue Waters

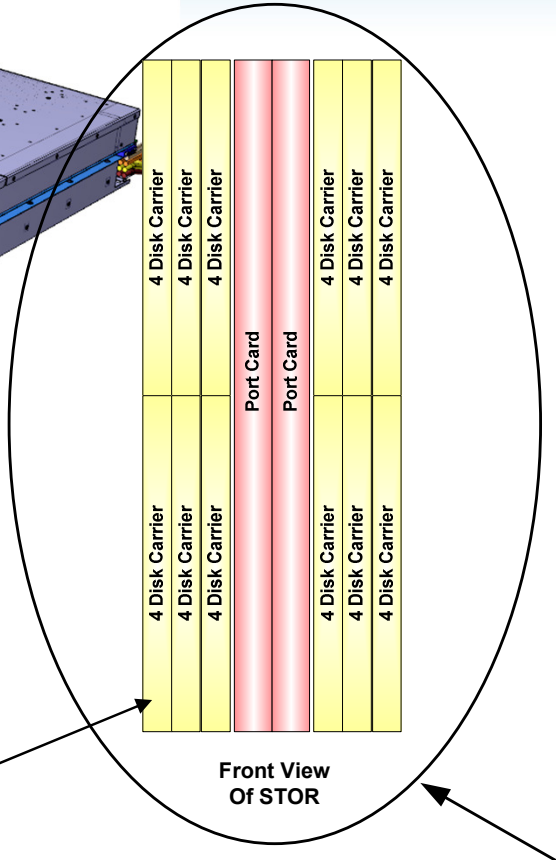
~10 PF Peak
~1 PF sustained
>300,000 cores
>1 PB of memory
>25 PB of disk storage
500 PB of archival storage
≥100 Gbps connectivity

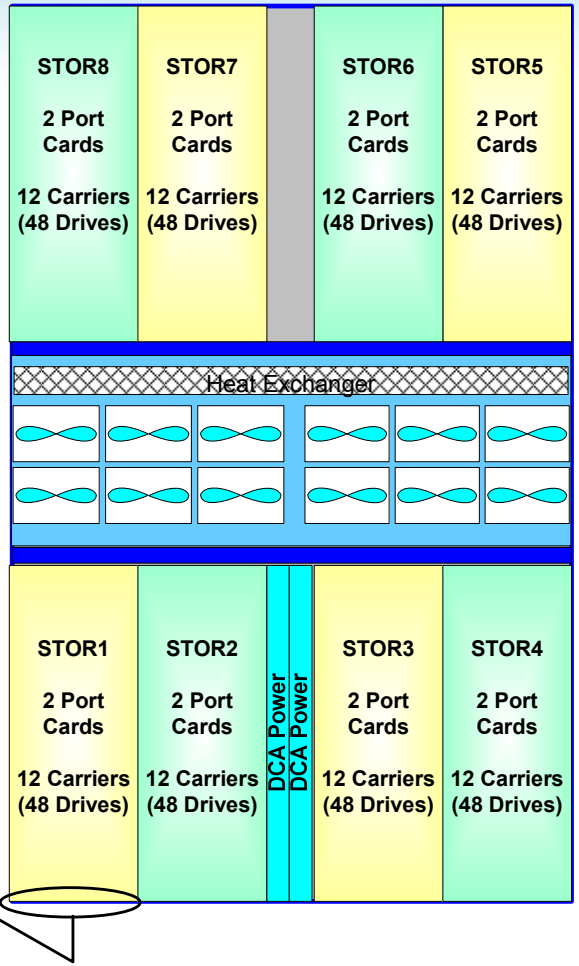
Gavetas de Armazenamento

- Até 384 disk-drives por gaveta
- Ligadas aos nós computacionais via PCIe
- Divididas em 8 grupos de armazenamento
  - 12 placas por grupo, 4 discos cada
  - 2 portas de E/S PCIe por grupo
- Sistema de arquivo: GPFS (IBM)



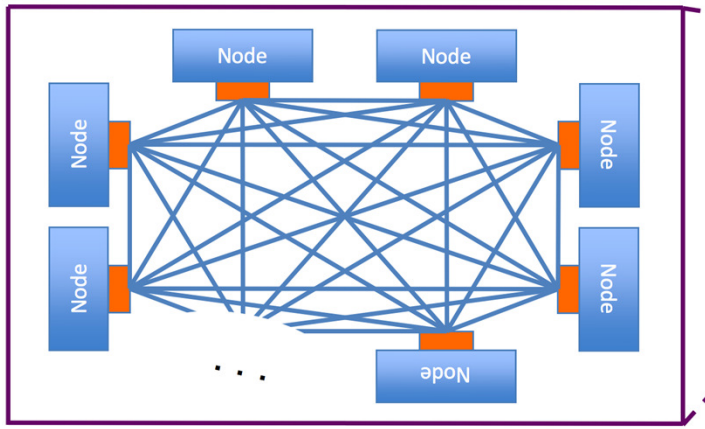



4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
Port Card
Port Card
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
Front View
Of STOR


STOR8
2 Port Cards
12 Carriers (48 Drives)
STOR7
2 Port Cards
12 Carriers (48 Drives)
STOR6
2 Port Cards
12 Carriers (48 Drives)
STOR5
2 Port Cards
12 Carriers (48 Drives)
Heat Exchanger
STOR1
2 Port Cards
12 Carriers (48 Drives)
STOR2
2 Port Cards
12 Carriers (48 Drives)
DCA Power
DCA Power
STOR3
2 Port Cards
12 Carriers (48 Drives)
STOR4
2 Port Cards
12 Carriers (48 Drives)

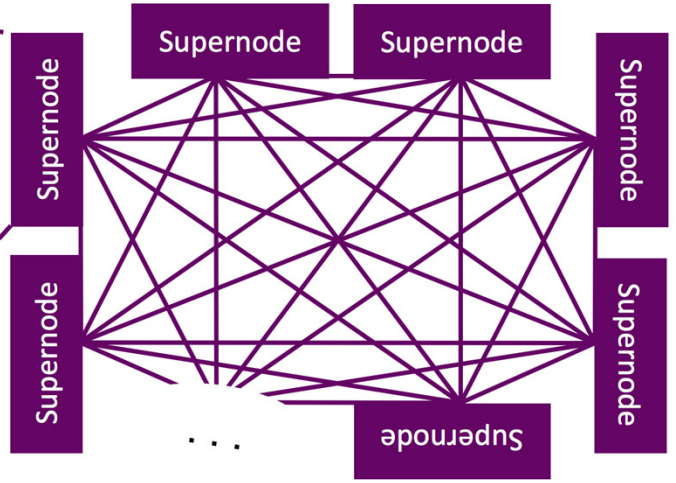
Rede de Interconexão: 2 Níveis

1 Supernó: 32 nós, em 4 gavetas



Conexões diretas entre nós:
Links LL: mesma gaveta (7)
Links LR: entre gavetas (24)

Config. Máxima: 513 Supernós

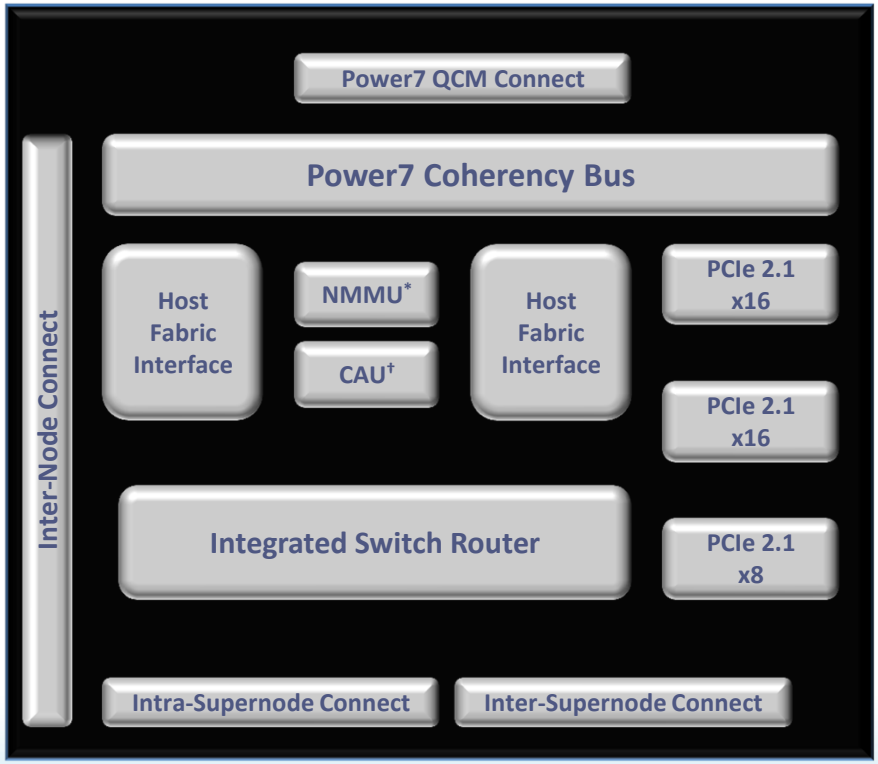


Conexões diretas entre Supernós:
Links D (32 nós × 16 links/nó = 512)

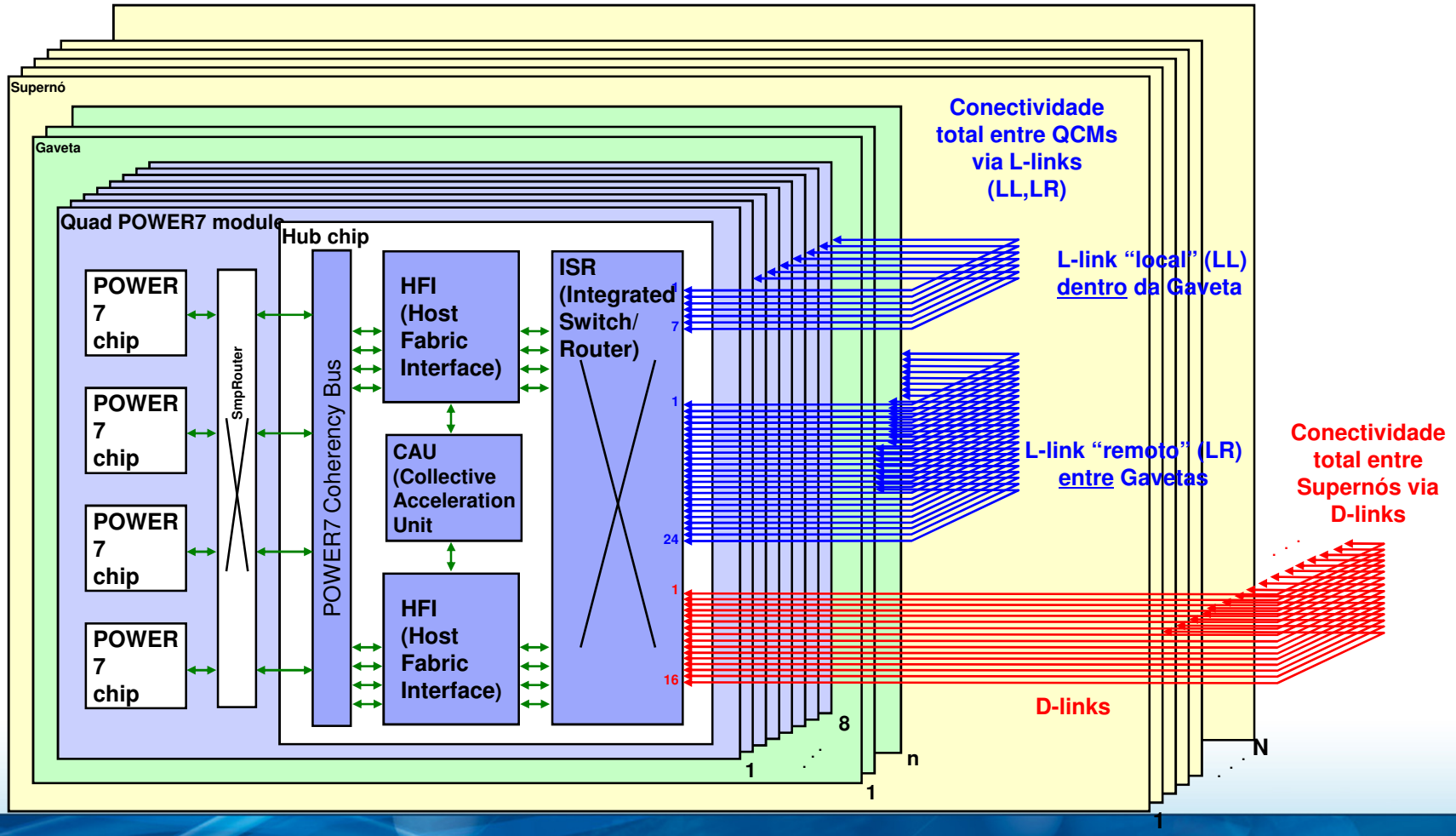
- **Roteamento entre dois nós quaisquer da máquina (pior caso):**
  - **Modo Direto: L – D – L**
  - **Modo Indireto: L – D – L – D – L**

Rede de Interconexão Interna

- Componente Básico: Hub = Torrent



Supernó: 4 Gavetas (1024 Núcleos, 32 TFlop/s)

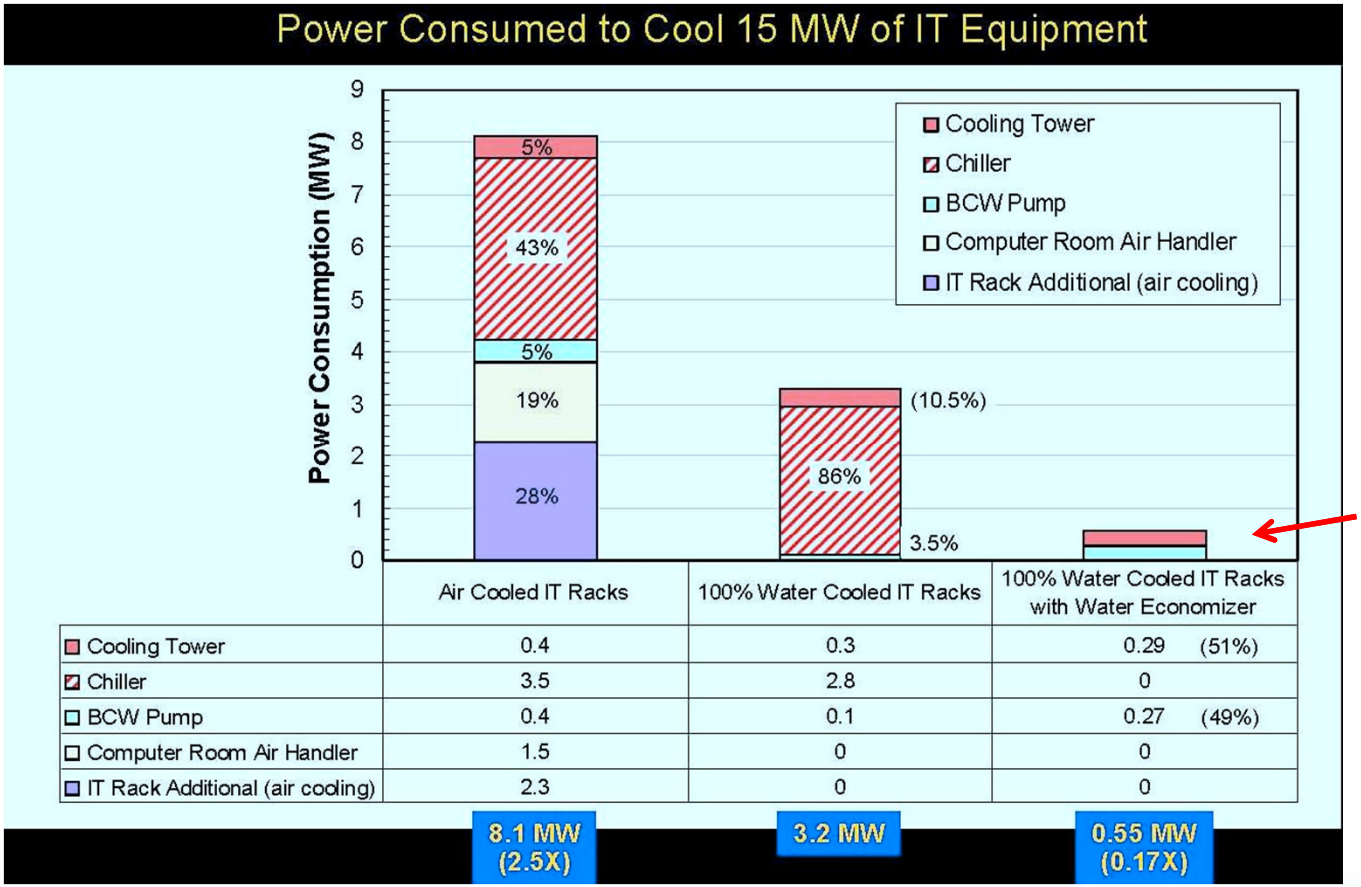


Prédio: "Petascale Computing Facility"

Características:

- Área total: 8.000 m²
- Sala de máq.: 2.700 m²
- Certific. Leed Gold
- PUE: 1.1~1.2
- Próxima à sub-estação de energia local
- 24 MW disponíveis
- 15 MW p/ consumo do Blue Waters (refrig. água; com ar, seriam 18 MW)
- Duas torres locais para resfriamento de água no inverno: ~6 meses
- Alta segurança física

Energia para Refrigeração



Inclui o uso de duas torres de resfriamento ao lado do prédio

Verão **Inverno ≈ 5~6 meses**

Software e Simuladores

- Software Previsto
  - Bibliotecas numéricas (ESSL, PESSL, PetSc, etc.)
  - Bibliotecas de comunicação (MPI, OpenMP, ARMCI/Global Arrays, LAPI, OpenSHMEM, etc)
  - Compiladores (Fortran, C/C++, UPC, CoArrayFortran)
- Simuladores Disponíveis
  - Mambo (IBM) – chip Power7
  - MARS (IBM) – rede de interconexão
  - BigSim (Illinois) – sistema completo

Aplicações em Desenvolvimento

- Suporte: NSF (apenas para viagens)
- Programa PRAC - Petascale-Resource Allocation
- Requisito: bom desempenho em máquinas atuais
- Prêmio:
  - Possibilidade de acesso ao Blue Waters
  - Apoio de pessoal especializado do NCSA
- Competição anual:
  - 18 grupos aprovados em 2008/2009
  - 5 grupos aprovados em 2010

Aplicações da Primeira Fase

- Ciências Biológicas
  1. Computational Microscope
  2. Petascale Simulations of Complex Biological Behavior in Fluctuating Environments
- Engenharia
  3. Petascale Computations for Complex Turbulent Flows
- Geo-Ciências
  4. Petascale Research in Earthquake System Science on Blue Waters
  5. Enabling Large-Scale, High-Resolution, and Real-Time Earthquake Simulations on Petascale Parallel Computers
  6. Testing Hypotheses about Climate Prediction at Unprecedented Resolutions on the Blue Waters System

Aplicações da Primeira Fase

- Geo-Ciências (cont.)
  7. Understanding Tornados and Their Parent Supercells Through Ultra-High Resolution Simulation/Analysis
- Ciências Astronômicas
  8. Computational Relativity and Gravitation at Petascale: Simulating and Visualizing Astrophysically Realistic Compact Binaries
  9. Enabling Science at the Petascale: From Binary Systems and Stellar Core Collapse to Gamma-Ray Bursts
  10. Formation of the First Galaxies: Predictions for the Next Generation of Observatories
  11. Peta-Cosmology: Galaxy Formation and Virtual Astronomy
  12. Petascale Simulation of Turbulent Stellar Hydrodynamics

Aplicações da Primeira Fase

- Química:
  13. Computational Chemistry at the Petascale
  14. Super Instruction Architecture for Petascale Computing
- Ciência de Materiais:
  15. Breakthrough Petascale Quantum Monte Carlo Calculations
  16. Electronic Properties of Strongly Correlated Systems Using Petascale Computing
- Física:
  17. Lattice QCD on Blue Waters
- Ciências Sociais e Econômicas
  18. Simulation of Contagion on Very Large Social Networks with Blue Waters

Distribuição Geográfica dos Grupos

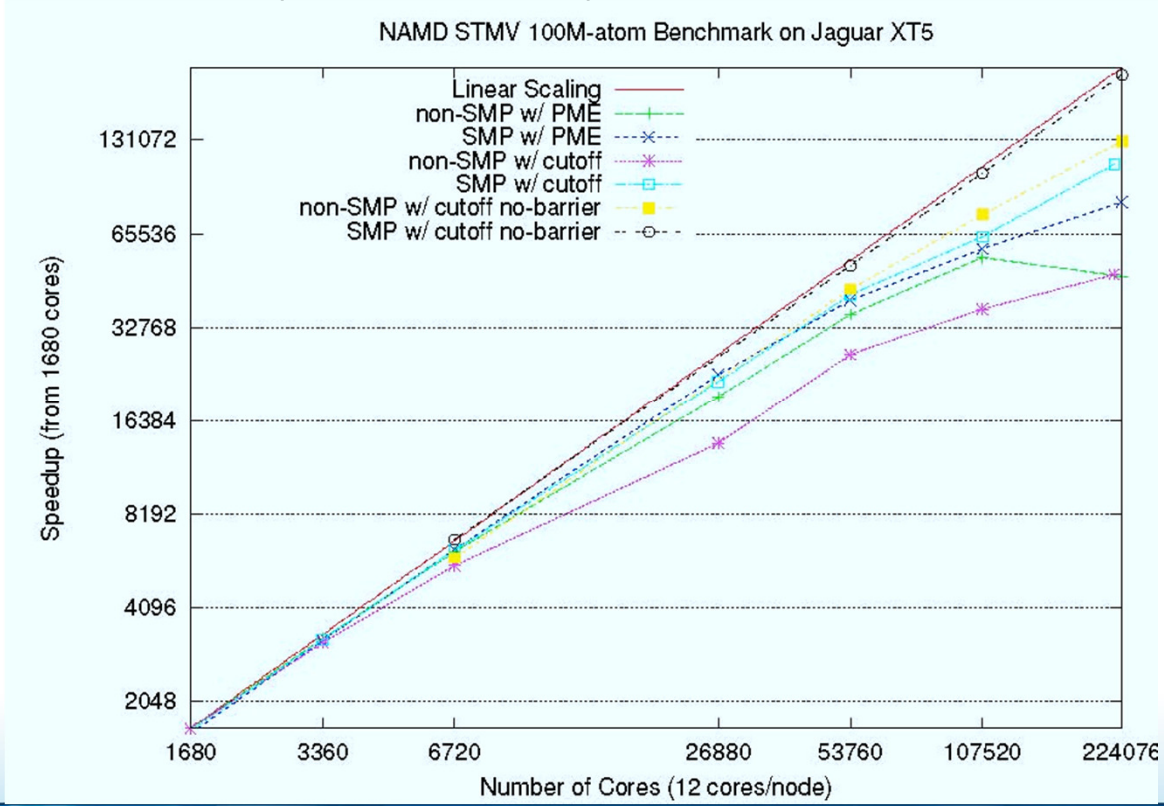
Principais Características das Aplicações

- Linguagens
  - Fortran/C/C++ na maioria
- Modelos de programação
  - Atual: MPI puro, MPI+OpenMP, Charm++
  - Em estudos: MPI+OpenMP, Charm++, UPC, Global Arrays, OpenSHMEM
- Desempenho em sistemas atuais
  - Serial: 6% a 32% do pico teórico
  - Escalabilidade: testes com até (a) P=130K no BlueGene/P, (b) P=220K no Cray-Jaguar

Aplicações e o Desafio da Escalabilidade (1)

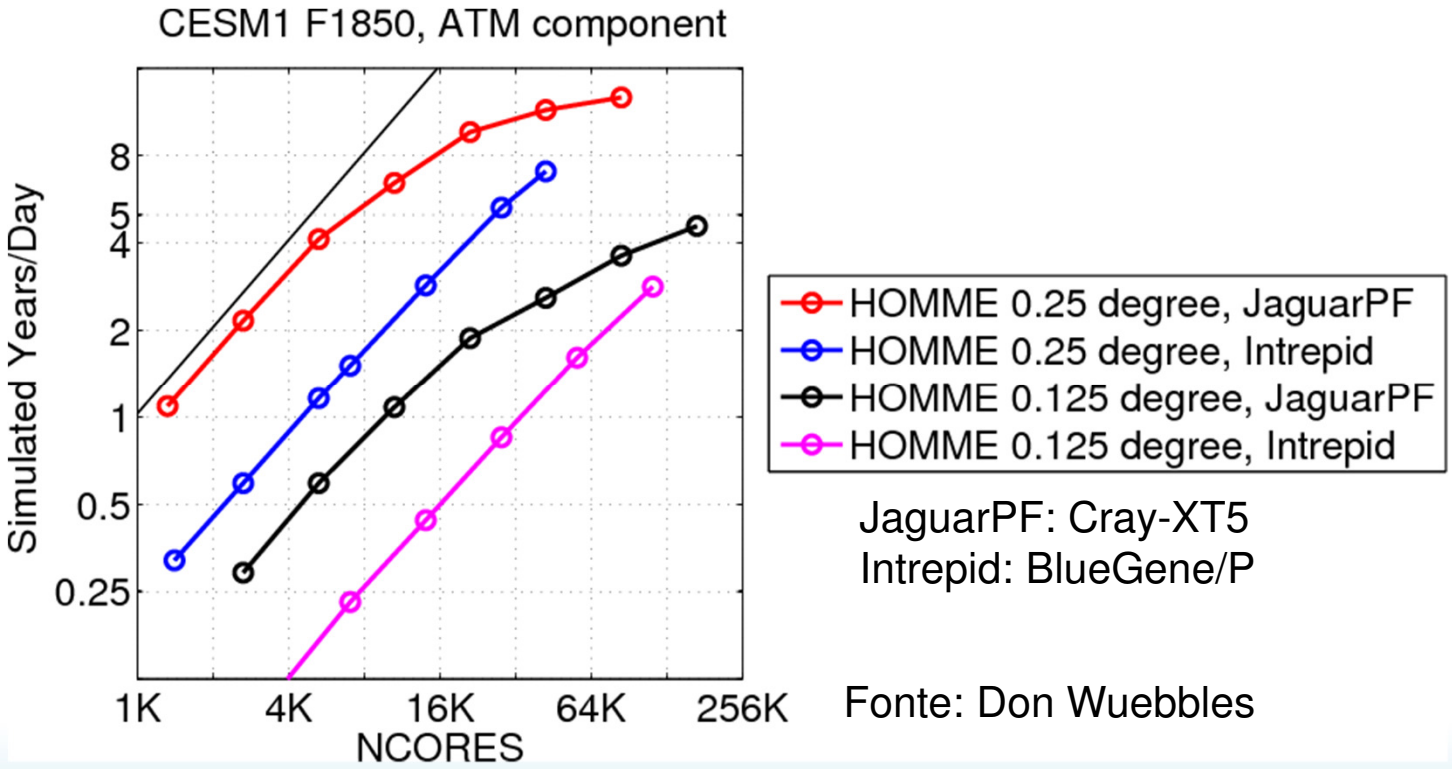
NAMD (Dinâmica Molecular, baseado em Charm++):

- Suporte para SMP melhora o desempenho
- Fase PME (contendo FFT) tem escalabilidade ainda limitada



Aplicações e o Desafio da Escalabilidade (2)

CCSM/HOMME (Modelo Acoplado de Clima):



JaguarPF: Cray-XT5
Intrepid: BlueGene/P

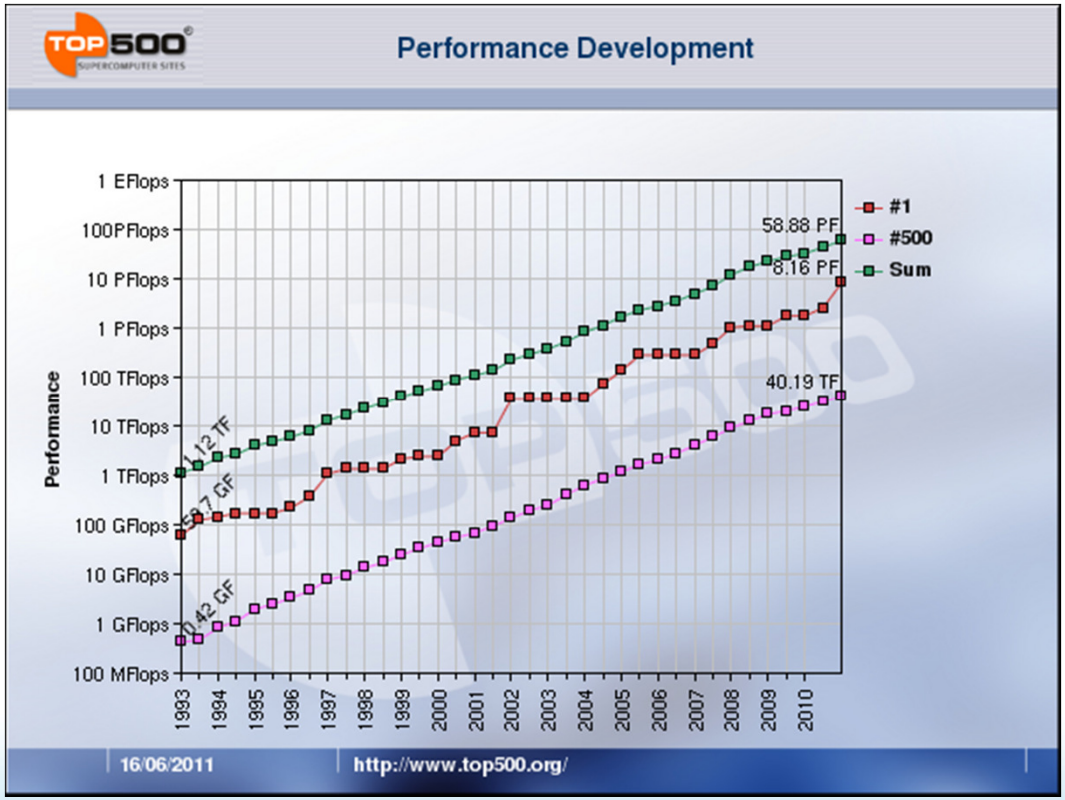
Fonte: Don Wuebbles

Tópicos:

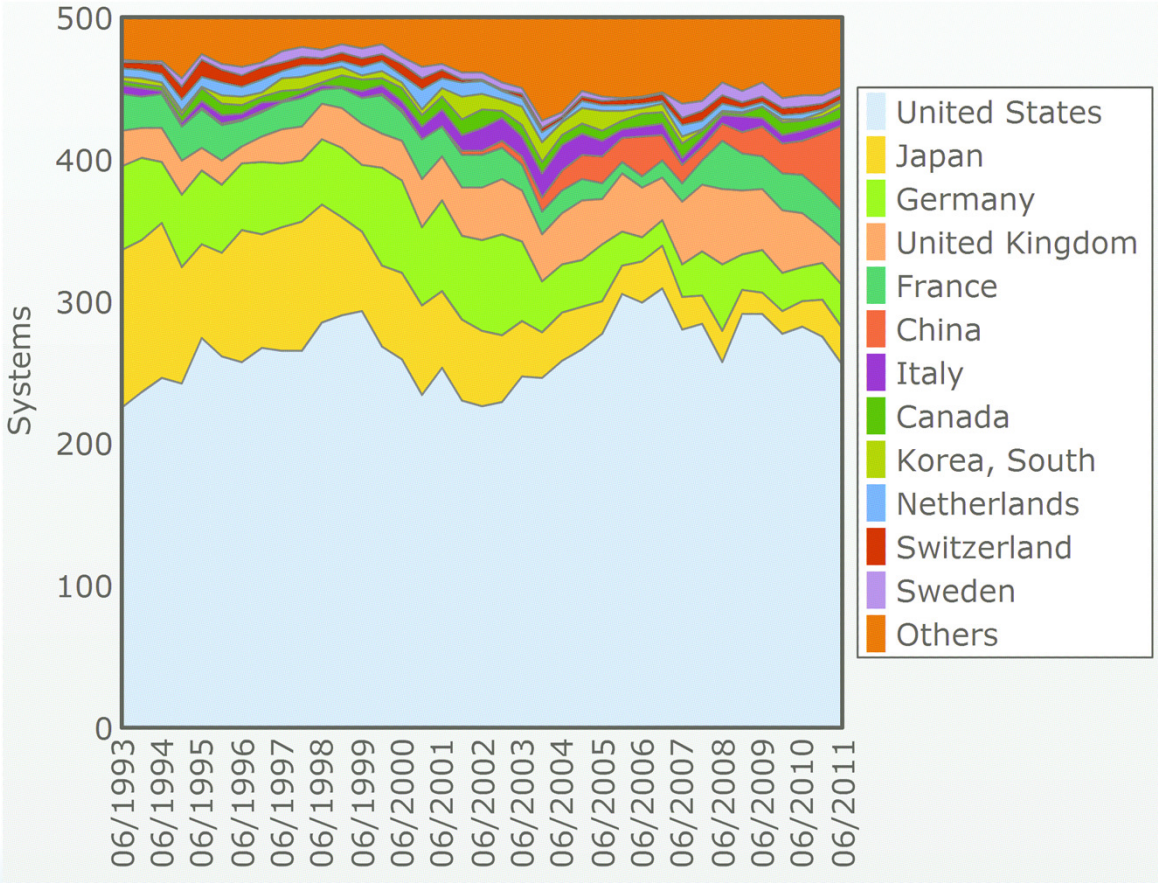
1. Sistema Blue Waters
   a) Hardware: CPU, Interconexão
   b) Software: Suporte, Applicações
2. Sistemas Petascale Atuais
3. Próxima Fronteira: Sistemas Exaflop
4. Outras Atividades Correntes
   a) Na Universidade de Illinois
   b) Em São José dos Campos

Sistemas Petascale Atuais

- http://www.top500.org: 2 listas anuais (Jun/Nov)



Sistemas Petascale Atuais – Por Países

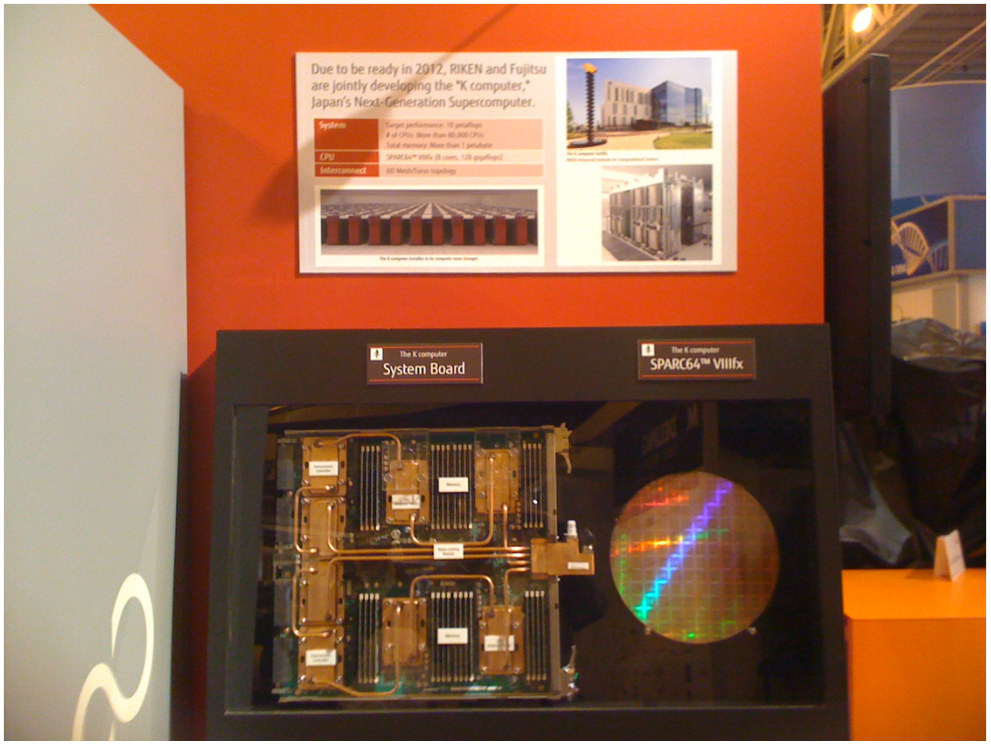


Top 5 – Resultados (Linpack)

Rank	Site	Computer/Year Vendor	Cores	R_{max} (Tflop/s)	R_{peak}	Power (KW)	% do pico
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	548352	8162.00	8773.63	9898.56	93.0%
2	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C / 2010 NUDT	186368	2566.00	4701.00	4040.00	54.6%
3	DOE/SC/Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60	75.5%
4	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU / 2010 Dawning	120640	1271.00	2984.30	2580.00	42.6%
5	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.61	52.1%

#1: K Computer (Fujitsu)

SC'10



K Computer

- Processador: Sparc64 @ 2.0 GHz
- Núcleos por processador: 8
- Total de núcleos: 548.352
- Desempenho planejado: 10 PFlop/s de pico (2012)
  - K = “Kei” (dez quadrilhões em Japones)
- Interconexão: Tofu – 6D mesh/torus
  - Artigo: IEEE Computer, Nov.2009, pg.36
- Localização: AICS Center - Kobe, Japão

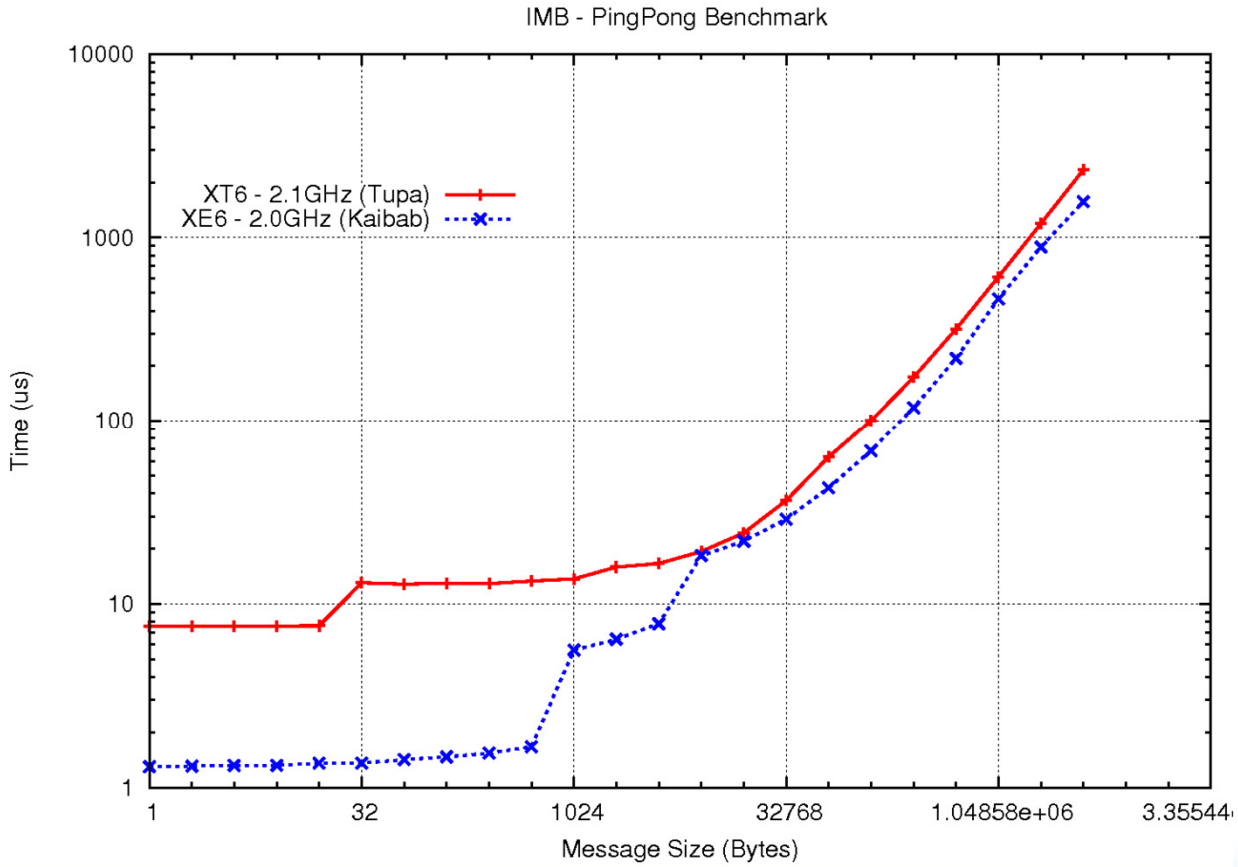
#3: Jaguar (Cray-XT5, EUA)

Atributo do Sistema	ORNL Jaguar (#3)	NCSA Blue Waters
Fabricante (Modelo)	Cray (XT5)	IBM (PERCS)
Processador	AMD Opteron	IBM Power7
Desempenho de Pico (PFlop/s)	2.3	≳10
Desempenho Efetivo (PFlop/s)	?	≳1
Número de Nucleos/Chip	6	8
Número Total de Núcleos	224,256	≳300,000
Total de Memória (TB)	299	≳1,200
Larg. Banda de Memória (PB/sec)	0.478	≳5
Total de Armazenamento em Disco (PB)	10	≳25
Taxa Efetiva p/de Disco (TB/sec)	0.24	≳1.5

#36: Tupã (Cray-XT6, Brasil)

- Processador: AMD@2.1GHz
- Núcleos por nó: 2 × 12 = 24
- Total de núcleos: ~30.500
- Desempenho: 258 TFlop/s de pico (no fim da instalação)
- Interconexão: torus/SeaStar (a ser trocada por Gemini)
- Localização: INPE / CPTEC (Cachoeira Paulista, SP)
- Aplicação: Previsões de Tempo e Clima

Tupã: Ganho Esperado com a Rede Gemini

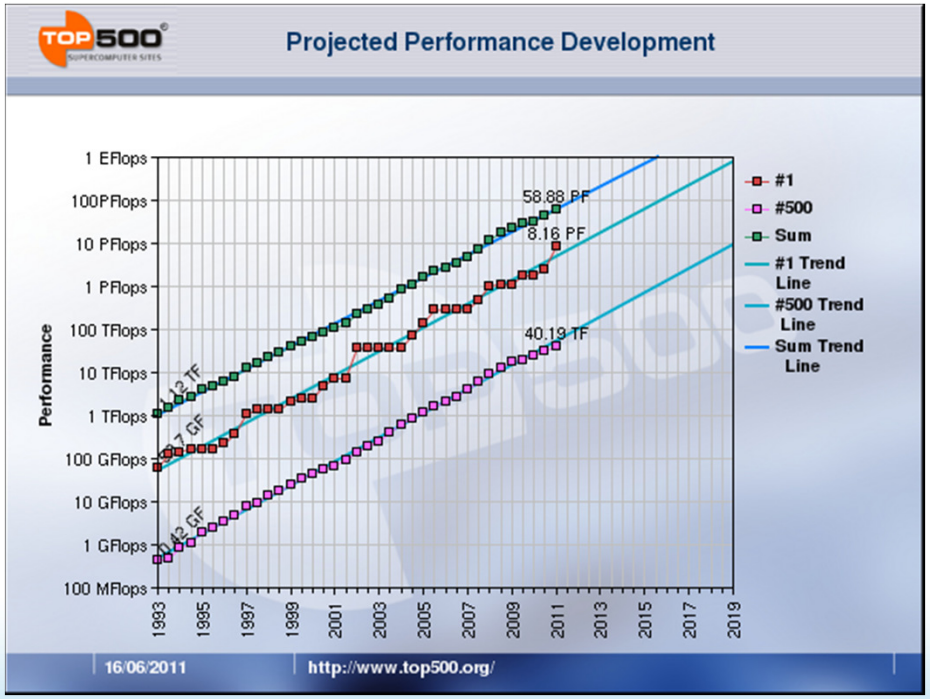


Tópicos:

1. Sistema Blue Waters
   a) Hardware: CPU, Interconexão
   b) Software: Suporte, Applicações
2. Sistemas Petascale Atuais
3. Próxima Fronteira: Sistemas Exaflop
4. Outras Atividades Correntes
   a) Na Universidade de Illinois
   b) Em São José dos Campos

Proxima Fronteira: Sistemas Exaflop

- 1 Exaflop = 10^{18} flops (um quinquilhão de flops)
- Previsão (pelo Top500): ~2019



Systems	2010	2018	Difference Today & 2018
System peak	**2 Pflop/s**	**1 Eflop/s**	**O(1000)**
Power	**6 MW**	**~20 MW**	
System memory	0.3 PB	32 - 64 PB	O(100)
Node performance	125 GF	1,2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s	2 - 4TB/s	O(100)
Node concurrency	12	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	225,000	O(billion)	O(10,000)
Storage	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO	0.2 TB	60 TB/s (how long to drain the machine)	O(100)
MTTI	days	O(1 day)	- O(10)

Fonte: Jack Dongarra

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

Fonte: Horst Simon

Sistemas Exaflop: Desafios

- Hardware:
  - Arquitetura de processador e memória
  - Consumo de energia
  - Tolerância a falhas, devido ao número de componentes
- Software:
  - Modelos de programação
    - MPI? MPI+X? X? Como controlar ~1 bilhão de threads?
  - Sistema Operacional
    - Sincronização, controle de jitter, etc.

Sistemas Exaflop: Atividades Correntes

- http://www.exascale.org
  - [Documents]: rica fonte de material
- Líderes:
  - Jack Dongarra & Pete Beckman
- Objetivo:

  "Build an international plan for coordinating research for the next generation open source software for scientific high-performance computing"

Tópicos:

1. Sistema Blue Waters
   a) Hardware: CPU, Interconexão
   b) Software: Suporte, Applicações
2. Sistemas Petascale Atuais
3. Próxima Fronteira: Sistemas Exaflop
4. Outras Atividades Correntes
   a) Na Universidade de Illinois
   b) Em São José dos Campos

University of Illinois at Urbana-Champaign

- Tradição em Proc. Alto Desempenho:
  - Série Illiac, incluindo o Illiac-IV (décadas de 60/70)
  - NCSA: criado em 1985 (Telnet, Mosaic, HDF, etc)
  - Computação com GPUs, GPU clusters

Prof.Hwo (ECE)

#3 no Green500 em Nov./2010

www.green500.org

Fonte: NCSA

University of Illinois at Urbana-Champaign

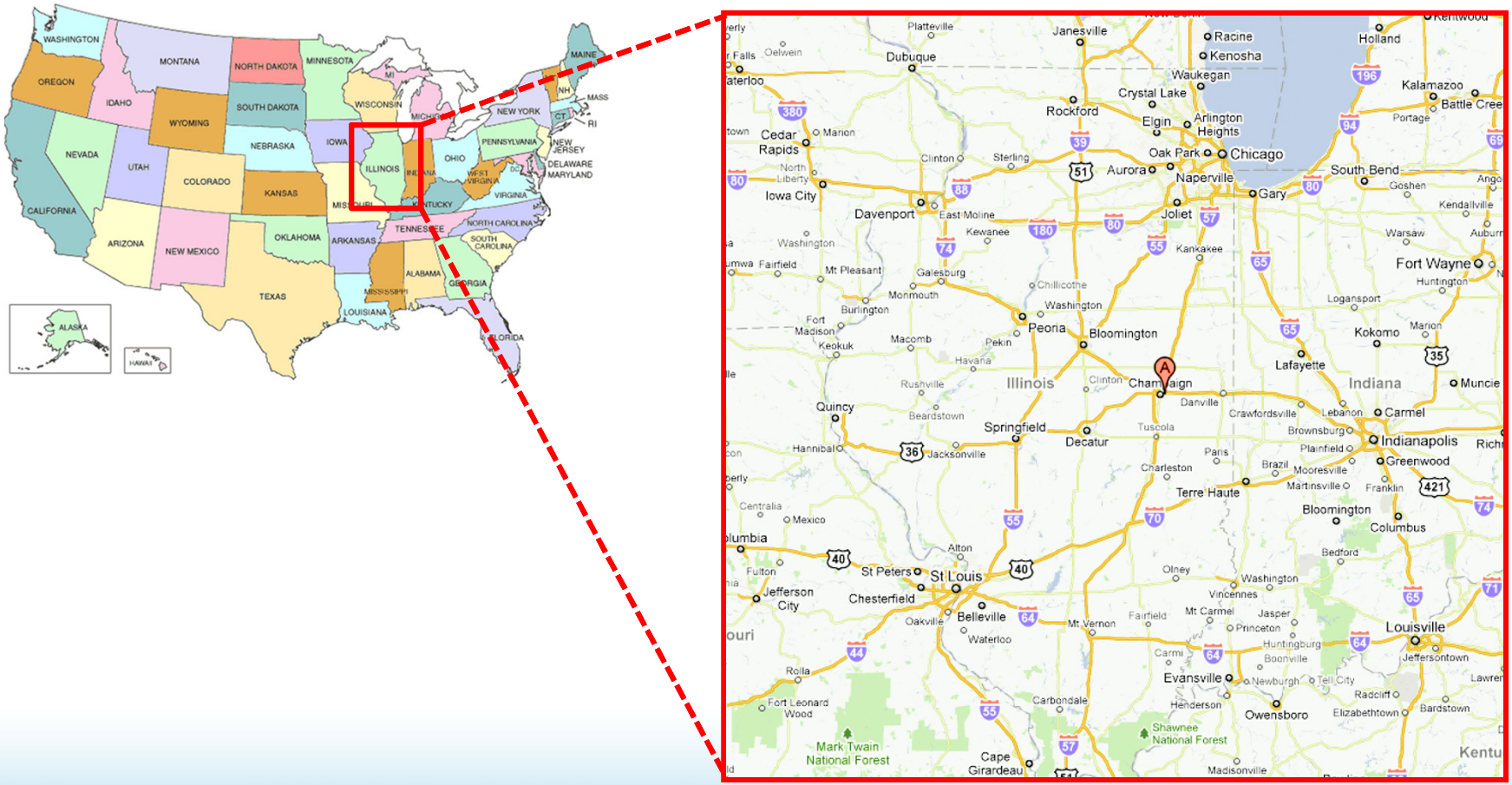
http://www.parallel.illinois.edu

Arquivo de palestras em vídeo

University of Illinois at Urbana-Champaign

- Atividades atuais em Proc. Alto Desempenho:
  - UPCRC (Prof. Snir): Univ. Par. Computing Research Center
    - financ. Intel/Microsoft, $10M em 5 anos
  - CCT – Cloud Computing Testbed: financ. HP/Intel/Yahoo!
  - IACAT – Inst. for Advanced Computing (Prof. Gropp)
  - PCI – Parallel Computing Institute (Profs. Gropp, Hwo)
  - Joint Lab. For Petascale Computing – colab. INRIA/Franca
  - CUDA Center of Excellence – pesq. em sistemas com GPU
  - OpenSPARC Center of Excellence – pesq. em arquitetura
  - Cursos: Graduação e Pós-Grad. (CS, ECE, CSE, …)
  - Etc, etc, etc

Illinois, Urbana-Champaign



Dias "Típicos" em Urbana

Inverno
8/Fev/2011: -4F (-20C)

Verão
11/Julho/2011: 114F (+45.5C)

S.J.Campos & PAD

- Cursos no INPE: Dr. Stephan Stephany
  - CAP-236: Computação Aplicada II (em parte)
  - CAP-372: Processamento de Alto Desempenho
- Curso no ITA: Dr. Jairo Panetta
  - CE-265: Processamento Paralelo
- Cursos na Unifesp (grad.): Prof. Álvaro Fazenda
  - PPD: Programação Concorrente e Distribuída
  - AD: Alto Desempenho

OBRIGADO!

- Dúvidas ?

cmendes@illinois.edu