UNIP

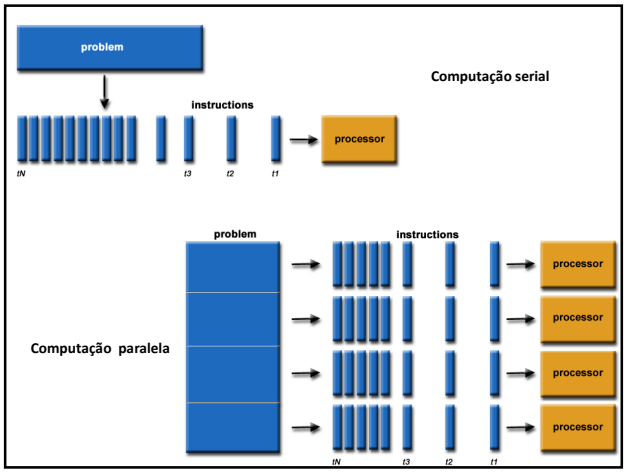
Sistemas Distribuídos

HPC – Introdução

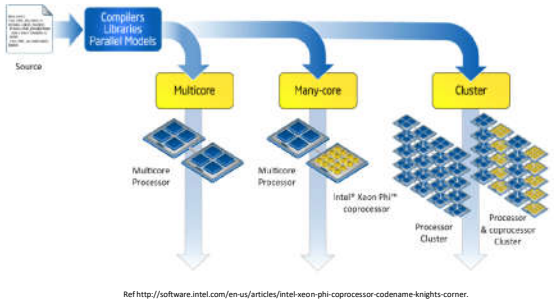
Luiz Carlos Magrini

Paralelismo

- Consiste em executar simultaneamente várias partes de uma mesma aplicação ;
- Tornou-se possível a partir do desenvolvimento de sistemas operacionais multi-tarefa, multi-thread e paralelos;
- Aplicações são executadas paralelamente:
  - Em um mesmo processador (pseudoparalelismo);
  - Em uma máquina multiprocessada;
  - Em um grupo de máquinas interligadas que se comportam como uma só máquina;
  - *GPGPU , isto é,* Unidade de Processamento Gráfico de Propósito Geral, ou **GPGPU** (General Purpose Graphics Processing Unit).



Multicore vs. Many core

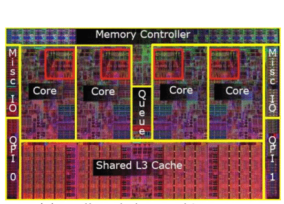


Ref http://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-codename-knights-corner.

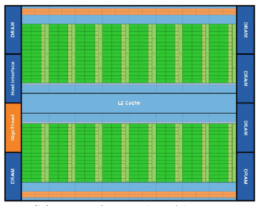
Computação Concorrente

- Existem múltiplas tarefas a serem executadas. (carros em movimento)
- Cada tarefa pode ser executada:
  - uma de cada vez em um único processador (uma única estrada ou pseudoparalelismo);
  - em paralelo em múltiplos processadores (várias pistas de uma estrada); ou,
  - em processadores distribuídos (estradas separadas);
  - Em GPGPUs, que possibilitam o processamento de várias threads em paralelo.

Arquitetura de CPU multi-core e de GPU many-core



(a) Intel's Nehalem Architecture

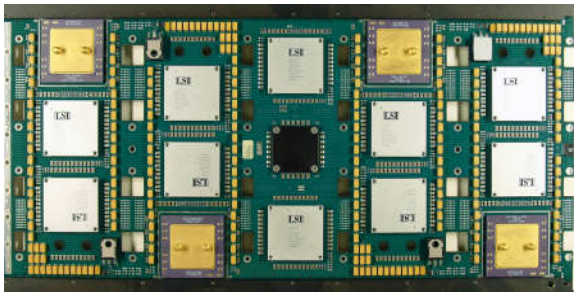


(b) NVIDIA's Fermi Architecture

Características

- Um programa concorrente contém dois ou mais processos que trabalham juntos para executar uma tarefa;
- Cada processo é um programa sequencial;
- Programa sequencial: corresponde a um única *thread* de controle;
- Programa concorrente: múltiplas *threads* de controle.

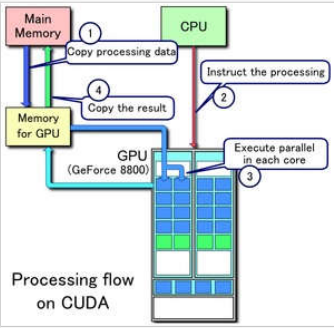
7



MotherBoard (Placa mãe) multi-CPU do supercomputador CRAY-2

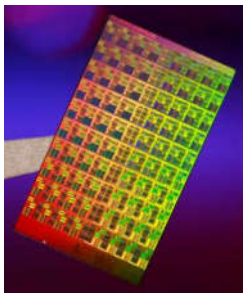
GPGPU

AMD Radeon RX 5700



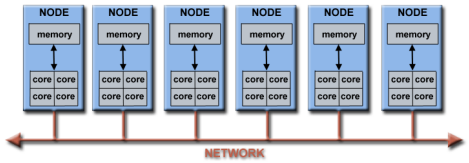
Exemplo de fluxo de processamento CUDA
1. Copiar dados de mem principal para GPU mem
2. CPU instrui o processo para GPU
3. GPU executar em paralelo em cada núcleo
4. Copie o resultado da GPU mem to main mem

Chip experimental Intel com 80 cores



SD – Fracamente acoplado

Sistema Distribuído (SD)



SD: Definição

- "Um *Sistema Distribuído (SD)* é uma coleção de processadores que **não compartilham memória** nem um ***clock***. Em vez disso, cada processador tem sua própria memória local, e os processadores se comunicam entre si por meio de várias redes de comunicação"

  *(SILBERSCHATZ, Abraham; et. al.**Sistemas Operacionais**. Rio de Janeiro: Campus,2000.)*

SD: Definição

- "Sistema no qual componentes de **hardware e/ou software**, localizados em computadores interligados em rede, se comunicam e coordenam suas ações apenas enviando mensagens entre si"
- *(COULOURIS, George; et. al.**Distributed Systems: Concepts and Design**.)*

SD: Definição

- **Sloman, 1987**

  "Um sistema de processamento distribuído é tal que, vários processadores e dispositivos de armazenamento de dados, comportando processos e/ou bases de dados, interagem cooperativamente para alcançar um objetivo comum. Os processos coordenam suas atividades e trocam informações por passagem de mensagens através de uma rede de comunicação**".**

SD: Definição

- *"Um sistema distribuído é um sistema formado por uma coleção de máquinas autônomas conectadas por redes de comunicação e equipadas com um software adequado para produzir um ambiente computacional integrado e consistente. Sistemas distribuídos possibilitam que as pessoas possam agir cooperativamente e coordenar suas atividades mais efetivamente e eficientemente*.

  *(JIA, Weijia; ZHOU, Wanlei. **"Distributed Network Systems – from concepts to implementations"**. Boston: Springer, 2005).*

SD: Definição

- Andrew **Tanenbaum:**

  "Coleção de computadores independentes que se apresenta ao usuário como um sistema único e consistente (coerente)" .

- **Coulouris**

  "Coleção de computadores autônomos interligados através de uma rede de computadores e equipados com software que permita o compartilhamento dos recursos do sistema: hardware, software e dados".

Compartilhamento de recursos

- "Recurso": termo abstrato e compreende tanto objetos de hardware como discos e impressoras quanto entidades de software como arquivos e banco de dados.
- Motivações:
  1 - Economia: compartilhamento de impressoras, supercomputadores, sistemas de armazenamento, etc. . .

  2 - Colaboração e troca de informações: arquivos, correio eletrônico, documentos, áudio e vídeo. Groupware, teleconferência, etc. . .

Computacão Distribuída

- **Computação Distribuída** é um método de processamento computacional na qual diferentes partes de um programa rodam simultaneamente em um ou mais computadores através de uma rede de computadores;
- É um tipo de **processamento paralelo**;
- **Sistema de processamento distribuído ou paralelo**: é um sistema que interliga vários nós de processamento (computadores individuais), não necessariamente homogêneos de maneira que um processo de grande consumo seja executado naquele nó "mais disponível", ou mesmo subdividido por vários nós.

SD: Computação Distribuída

- A **Computação Distribuída** fornece toda a infra-estrutura necessária para a construção e operação efetiva de aplicações distribuídas;
- Inclui todos os produtos necessários para permitir que essas aplicações sejam construídas, e possam ser executadas, em um **ambiente de rede heterogêneo**, ou em um **ambiente centralizado.**

SD x Paralelos

- **Acoplamento:**
  - Sistemas paralelos são **fortemente acoplados**: compartilham hardware ,ou se comunicam através de um barramento de alta velocidade;
  - Sistemas distribuídos são **fracamente acoplados**.
- **Previsibilidade:**
  - O comportamento de sistemas paralelos é mais previsível;
  - Os sistemas distribuídos são mais imprevisíveis devido ao uso da rede e a falhas.

Definição

- **Computação Distribuída** é um método de processamento computacional na qual diferentes partes de um programa rodam simultaneamente em um ou mais computadores através de uma rede de computadores.
- É um tipo de **processamento paralelo**.
- **Sistema de processamento distribuído ou paralelo**: é um sistema que interliga vários nós de processamento (computadores individuais), não necessariamente homogêneos de maneira que um processo de grande consumo de CPU seja executado no nó "mais disponível", ou mesmo subdividido por vários nós.

22

Definição

- **Paralelismo:** divisão de uma tarefa em sub-tarefas coordenadas e que são executadas simultaneamente em processadores distintos

23

Redes de Computadores x Sistemas Distribuídos

- **Redes de Computadores**: é uma coleção de computadores separados interconectados que trocam mensagens baseadas em um protocolo específico. Os computadores são endereçados pelo endereço IP, numa rede TCP/IP
- **Sistema Distribuído:** vários computadores em rede trabalhando juntos como um sistema.
  - A separação espacial dos computadores e aspectos de comunicação são escondidos dos usuários

24

Atrativos

- Velocidade de processamento;
- Compartilhamento de recursos;
- Confiabilidade;
- Custo/desempenho;
- Independência de fornecedor.

Características (1)

- **Compartilhamento de recursos:**
  - Compartilhamento de equipamentos e dados
    recursos de hardware: discos, impressoras, …
    recursos de software: arquivos, banco de dados, …
    outros recursos: poder de processamento, memória, largura de banda, …
  - Uso da arquitetura cliente-servidor ;
  - Servidores que agem como clientes e servidores.

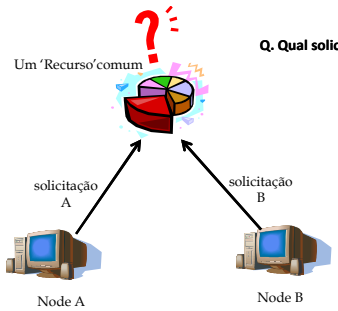
SD x Paralelos

- **Influência do Tempo:**
  - Sistemas distribuídos são bastante influenciados pelo tempo gasto na comunicação através da rede;
  - Em SD não há uma referência de tempo global geral;
  - Em sistemas paralelos o tempo gasto na troca de mensagens pode ser desconsiderado.
- **Controle:**
  - Em sistemas paralelos tem-se o controle de todos os recursos computacionais;
  - Os sistemas distribuídos tendem a empregar também recursos de terceiros.

Relógio do Sistema

- Em computadores, sequenciamento é tudo. O relógio do sistema sincroniza as tarefas de um computador, tais como carregamento dos dados na CPU antes da sua manipulação, etc.
- O relógio do sistema é um circuito que emite um fluxo continuo e preciso de pulsos altos e baixos que são exatamente da mesma duração;
- Um ciclo de clock é o tempo decorrido do início de um pulso alto até o início do próximo pulso;
- Caso diversos eventos irão acontecer em um ciclo de clock, o ciclo é subdividido colocando-se um circuito com um atraso conhecido, provendo dessa forma mais altos e mais baixos.

Problema!



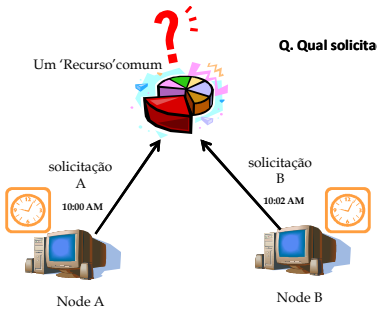
Solução

Um relógio global ??

Sincronização Global ?

Um Sistema Distribuído

Problema!



Solução

Relógio Individual?

Os clocks individuais são exatos, precisos?

Um clock poderia rodar mais rápido, ou mais devagar?

Um Sistema Distribuído

30

Clock

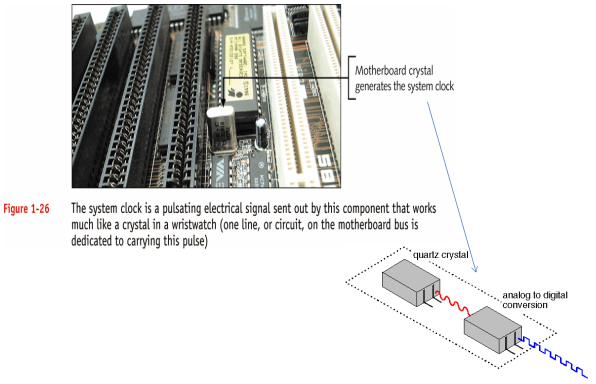


Figure 1-26 The system clock is a pulsating electrical signal sent out by this component that works much like a crystal in a wristwatch (one line, or circuit, on the motherboard bus is dedicated to carrying this pulse)

Problemas com os Clocks dos Computadores

- **Skew:** Desacordo com a leitura de dois clocks;
- **Drift:** Diferença na frequência gerada pelo oscilador do clock, que interfere na forma como duas máquinas contabilizam o tempo
  - Devido a diferenças físicas dos cristais, além de aquecimento, umidade, tensão, etc.
  - O drift acumulado ao longo do período que a máquina fica ligada pode levar a um skew significativo;
- Taxa de drift do Clock: Diferença na precisão entre um relógio de referencia e um clock físico;
  - Usualmente, 10^{-6} s/s;
  - Em relógios de elevada precisão: de 10^{-7} a 10^{-8} s/s.

Algoritmos Distribuídos

Os algoritmos distribuídos possuem as seguintes diferenças com relação aos centralizados:

1. Nenhum nó possui informação completa do estado do sistema;
2. Cada nó toma decisões baseado somente em informações locais;
3. A falha de um nó não inviabiliza a execução do algoritmo;
4. Não se pressupõe a existência de um relógio global.

SD: Funcionalidades

- Um SD deve prover:
  - Gerenciamento da comunicação interprocessos;
    - Perdas de mensagens
  - Sincronização de processos;
    - Diversos usuários simultâneos podem colocar em risco a integridade dos dados:
      - Problemas de exclusão mutua e sincronização: ex: processamento de transações em SGBD
      - Dados replicados: consistência da informação?
  - Tratamento de *deadlocks*;
  - Tratamento de outras situações não encontradas em sistemas centralizados .

SD: Definição

- Andrew **Tanenbaum:**

  "Coleção de computadores independentes que se apresenta ao usuário como um sistema único e consistente (coerente)" .

- **Coulouris**

  "Coleção de computadores autônomos interligados através de uma rede de computadores e equipados com software que permita o compartilhamento dos recursos do sistema: hardware, software e dados".

Definição

- **Paralelismo**: Divisão de uma tarefa em sub-tarefas coordenadas, e que são executadas simultaneamente em processadores distintos.

36

Interoperabilidade, Portabilidade e Extensibilidade

- **Interoperabilidade** caracteriza até que ponto duas implementações de sistemas ou componentes de fornecedores diferentes devem coexistir e trabalhar em conjunto, especificados por um padrão comum;
- **Portabilidade** caracteriza até que ponto uma aplicação desenvolvida para um sistema distribuído A pode ser executada, sem modificação, em um sistema B;
- **Extensibilidade ou escalabilidade:** define a capacidade de se adicionar novos componentes ou substituir componentes existentes sem afetar os que continuam no mesmo lugar.

Atrativos

- Velocidade de processamento;
- Compartilhamento de recursos;
- Confiabilidade;
- Custo/desempenho;
- Independência de fornecedor.

Desvantagens dos SD

- Falta de softwares adequados;
- Falhas e saturação da rede de comunicação podem eliminar as vantagens de SD;
- A segurança pode ser comprometida de maneira relativamente fácil;
- Acesso a dados e recursos reservados a aplicações desconhecidas ou de terceiros;
- Maior consumo de energia;
- Refrigeração, quando montados em racks.

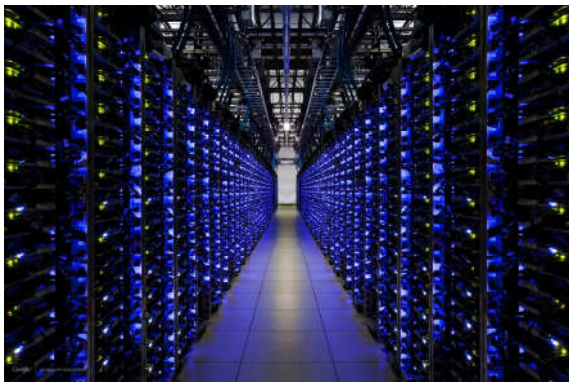
Sistemas Distribuídos

- Vantagens de Sistemas Distribuídos em relação a Sistemas Centralizados:
  - Preço: Hardware de baixo valor agregados;
  - Velocidade: é possível construir sistemas com valor agregado muito maior;
  - Distribuição física: algumas aplicações são essencialmente distribuídas (e.g., correio eletrônico);
  - Confiabilidade: se uma máquina quebra, outras podem guardar backup;
  - Disponibilidade: se uma máquina sai do ar, é possível utilizar outra;
  - Crescimento incremental: podemos acrescentar (ou retirar) recursos aos poucos .

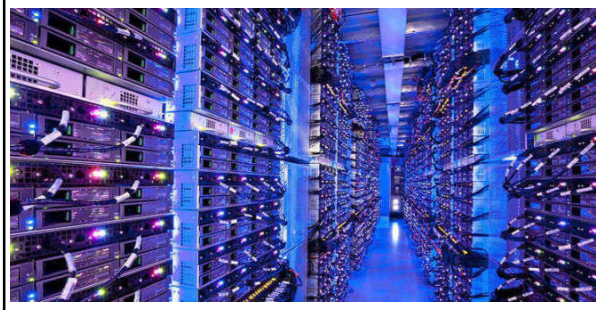
Conceitos Equivocados

- A rede de comunicação é confiável;
- A rede de comunicação é segura;
- A rede de comunicação é homogênea;
- A topologia da rede de comunicação não se altera;
- A latência da rede de comunicação é zero;
- A largura de banda da rede de comunicação é infinita;
- O custo de transporte das mensagens é zero;
- Há um único administrador;
- Há um conhecimento inerente compartilhado.

Cluster Google



Cluster Microsoft



Cloud Amazon

Cloud IBM

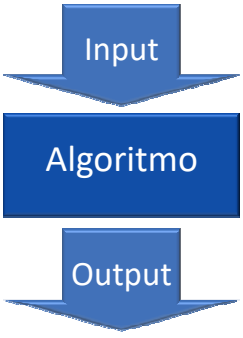
SD x Paralelos

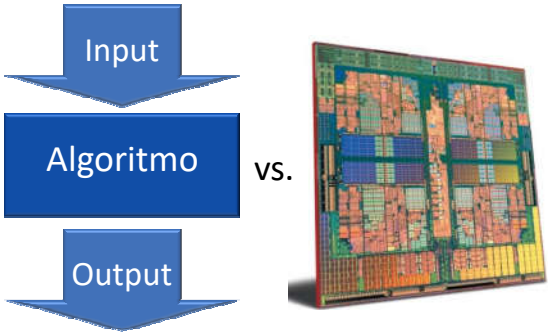
- **Acoplamento:**
  - Sistemas paralelos são fortemente acoplados: compartilham hardware ,ou se comunicam através de um barramento de alta velocidade;
  - Sistemas distribuídos são fracamente acoplados.
- **Previsibilidade:**
  - O comportamento de sistemas paralelos é mais previsível;
  - Os sistemas distribuídos são mais imprevisíveis devido ao uso da rede e a falhas.

SD x Paralelos

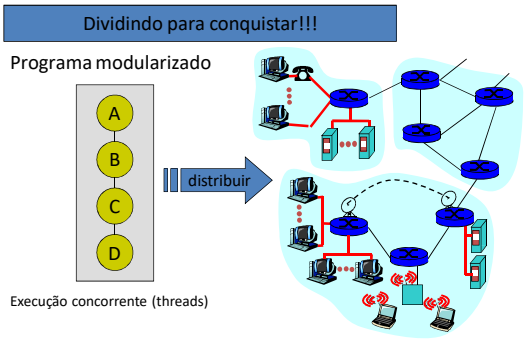
- **Influência do Tempo:**
  - Sistemas distribuídos são bastante influenciados pelo tempo gasto na comunicação através da rede;
  - Em SD não há uma referência de tempo global geral;
  - Em sistemas paralelos o tempo gasto na troca de mensagens pode ser desconsiderado.
- **Controle:**
  - Em sistemas paralelos tem-se o controle de todos os recursos computacionais;
  - Os sistemas distribuídos tendem a empregar também recursos de terceiros.

Execução sequencial convencional





Sistemas Distribuídos



Comunicação

- Os processos em um programa concorrente trabalham juntos comunicando-se entre eles;
- A comunicação pode ser realizada através de:
  - Variáveis compartilhadas (*shared memory*)
  - Troca de mensagens (*messages*)
- Independentemente da forma de comunicação, os processos precisam se sincronizar.

52

Algoritmos Distribuídos

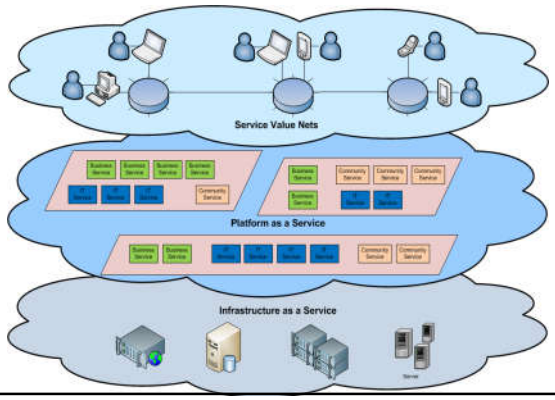
- Algoritmos que foram desenvolvidos para serem executados em muitos processadores "distribuídos" em uma grande área geográfica;
- Atualmente, o termo cobre não só algoritmos que são executados em redes locais, mas também em multiprocessadores de memória compartilhada (*shared memory*).

53

O que é Cloud Computing (ou computação em nuvem)

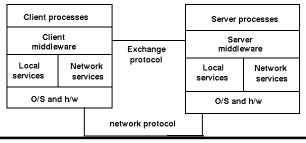
- Coleção de recursos computacionais compartilhados configuráveis;
- Acesso via rede, sob demanda;
- Entrega da computação como um serviço ao invés de um produto;
- Provisionado por um Provedor de Serviços.

Arquitetura Cloud



Middleware

- Middleware é uma classe de tecnologias de software projetadas para ajudar gerenciar a complexidade e heterogeneidades inerentes em sistemas distribuídos;
- É definida como uma camada de software acima do sistema operacional, mas que fica abaixo do programa aplicativo, e que provê uma camada de abstração comum a todo os sistema distribuído.

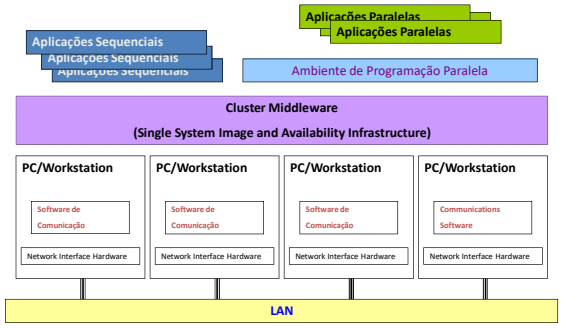


Bakken 2001

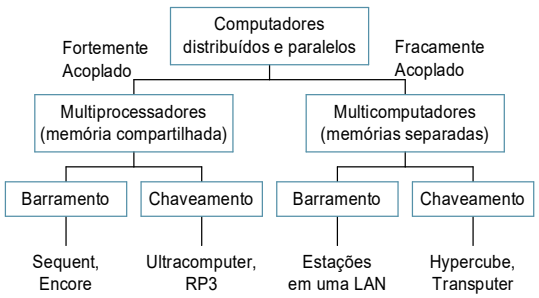
Middleware

- Conjunto de ferramentas que proporcionam um mecanismo uniforme e padronizado para acesso aos recursos do sistema em todas as plataformas;
- Possibilita que os programadores desenvolvam aplicações que se comportam de uma mesma forma, qualquer que seja a plataforma onde irá ser processado;
- Possibilita que os programadores utilizem um mesmo método de acesso aos dados em qualquer ambiente operacional.

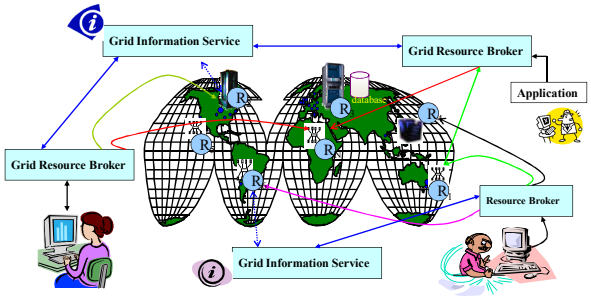
Cluster Computacional - Arquitetura



Hardware em SD



Grid Computacional Típico



Supercomputação ou HPC

- O que é Computação de Alto Desempenho?
  - Uso de computadores poderosos para resolver os maiores e mais complexos problemas.
- A Supercomputação também é conhecida como:
  - Processamento de Alto Desempenho (PAD)
  - High Performance Computing (HPC)
  - High Performance Technical Computing (HTPC)

HPA - Cluster de Alta Disponibilidade

- É construído com a intenção de fornecer um ambiente seguro contra falhas (*fail safe*) utilizando-se da redundância de componentes (*hardware, software, serviços de rede ou de interconectividade ou interoperabilidade*).
- Ou seja, fornecer um ambiente computacional onde a falha de um ou mais componentes não irá afetar significativamente a disponibilidade do ambiente de computação ou aplicações que estejam sendo usadas.

HPC - Cluster com Alto Desempenho Computacional

- É projetado para fornecer maior poder de computação para a solução de um problema.
- Está relacionado com aplicações científicas, de simulação ou de manipulação de imagens.
- O usuário interage com um nó específico para iniciar ou escalonar uma atividade que deverá ser executada.
- A aplicação, juntamente com as funções internas do *cluster*, irá determinar como a atividade será dividida e enviada para cada elemento que compõe o ambiente computacional, buscando extrair uma maior vantagem dos recursos disponíveis.

Cluster de Balanceamento de Carga

- É utilizado para fornecer uma interface simplificada para um conjunto de recursos que podem aumentar ou diminuir no balanceamento de carga com o passar do tempo e conforme a necessidade por processamento do cliente.
- Neste tipo de *cluster*, estão implícitos os conceitos da alta disponibilidade (com a redundância de componentes) e de alto desempenho de computação (com a distribuição das tarefas completas pelos vários componentes replicados).

Ambiente Computacional (cluster)

- **Evolução da capacidade de processamento:**

~ 1980: $1x10^6$ (MFLOP/s):
- **Processamento escalar;**

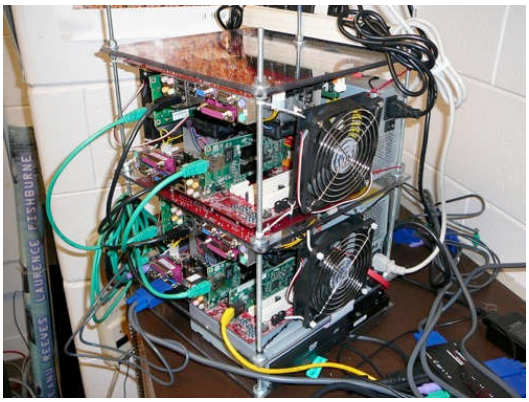
~ 1990: $1x10^9$ (GFLOP/s):
- **Processamento vetorial;**
- **Particionamento de dados;**

~ 2000: $1x10^{12}$ (TFLOP/s):
- **Processamento distribuído;**
- **Troca de mensagens;**
- **Decomposição de domínio;**

~ 2010 $1x10^{15}$ (PFLOP/s) (Jack Dongarra):
- **Grid Computer (computação em grade);**
- **...**

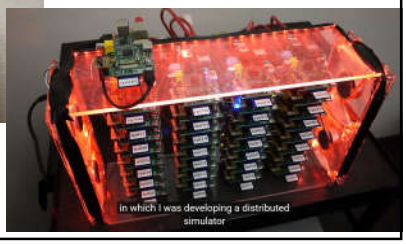
2 - Ambiente Computacional (hardware)

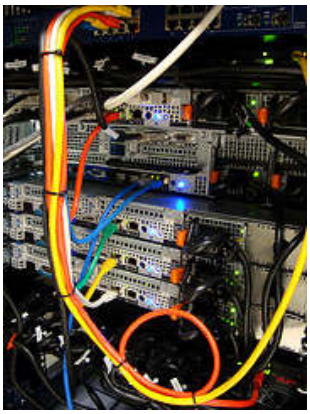
- **O hardware que forma o cluster pode ser:**
  - **Multiprocessado: arquitetura com memória distribuída inter nodal e compartilhada intra nodal;**
  - **Monoprocessado: arquitetura com memória distribuída;**
  - **Redes de interconexão mais adotadas:**
    - **Myrinet;**
    - **Fast-Ethernet;**
    - **Gigabit-Ethernet.**

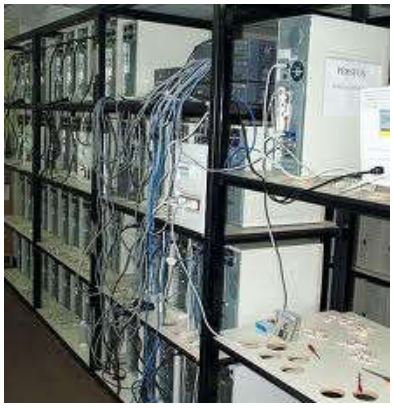


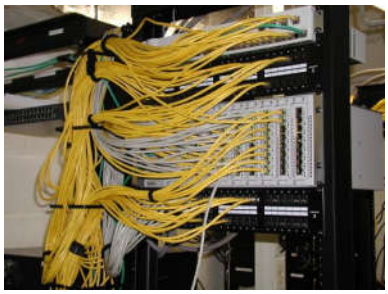
Raspberry Pi Cluster

33 nodes

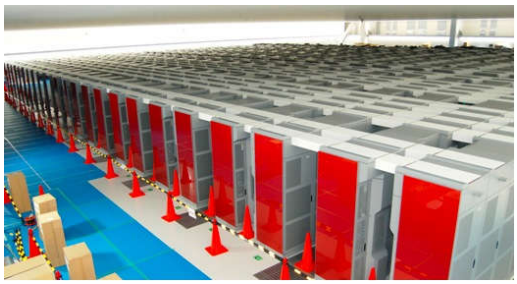








Fujitsu's 10.51 petaflop K supercomputer is fastest in the world



864 racks, 88.128 processadores SPARC64VIIIfx

- O computador mais rápido do mundo pelo quarto ano consecutivo, o Tianhe-2 fica localizado no National Supercomputing Center, na cidade de Tianjin.
- O Tianhe-2 possui 32 mil processadores Intel Xeon E5-2692 de 12 núcleos, que funcionam a 2,2 GHz.
- Ele ainda traz 48 mil coprocessadores Xeon Phi da Intel que são ligados a uma porta PCI Express, cada um com mais de 50 núcleos de processamento.
- No total o Tianhe-2 possui 3,12 milhões de núcleos de processamento e 1 petabytes de RAM.

Tianhe-2 (ou Via Láctea 2 - tradução)

TOP 10 Sites for June 2019

For more information about the sites and systems in the list, click on the links or view the complete list.

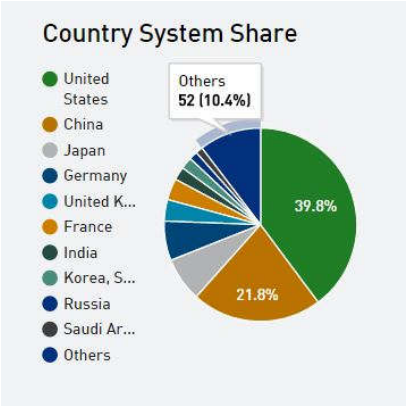
1-100 | 101-200 | 201-300 | 301-400 | 401-500

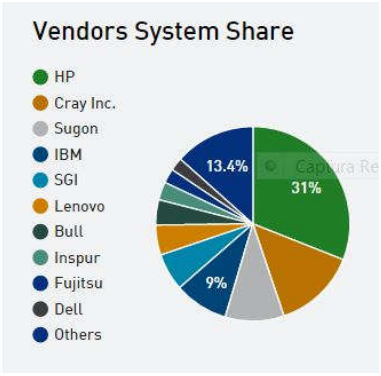
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
2	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
3	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 ,	4,981,760	61,444.5	100,678.7	18,482

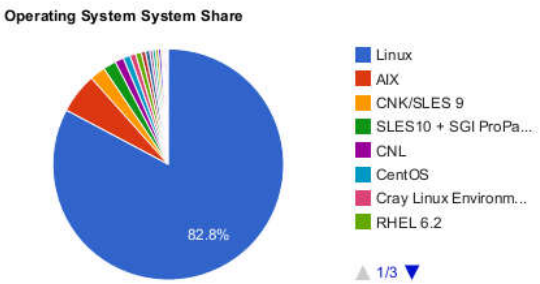
Brasil

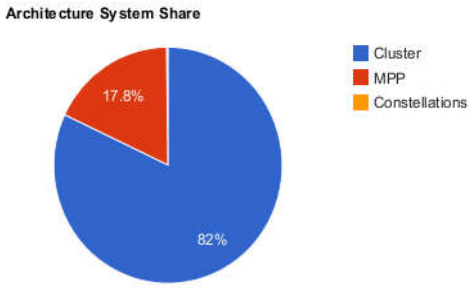
List	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Nov 2011	2	0.4	269730	330444.8	37184

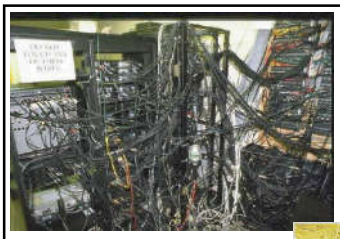
Posição 49 – INPE com CRAY
Posição 290 – COPPE com Sun











Cluster da Unicamp usa PlayStation 3

- Formado por 12 PlayStation 3, com disco rígido de 60 GB cada um, rodando Linux;
- Cada PlayStation tem um processador dual-core PowerPC, da IBM, que controla os seis processadores Cell contendo seis processadores por máquina – totalizando 72 processadores;
- Funcionam 24 horas por dia, sete dias por semana;
- São realizadas simulações que ajudam no estudo da interação de anestésicos locais utilizados em odontologia com membranas celulares (bioinformática).

Grifo04 - Petrobras

Grifo04 – Petrobras: 1088 GPUs , com 487 mil núcleos de processamento matemático, 17 TB de memória RAM (incluindo 3TB de GRAM) e 544 servidores com interface de rede com 20 Gigabits/segundo. Custo R$ 17 milhões.

Os três maiores computadores do Brasil:

1º.Grifo04 (Petrobras)	**251,5**	**68º.**
2º.Tupã (INPE)	214,2	79º.
3º.Galileu (UFRJ)	64,63	457.

Titan

Powered by a mixture of CPUs and GPUs, Titan is home to 18,688 nodes, each of which contains an AMD 16-core Opteron and a NVIDIA Tesla K20X GPU accelerator. Fica localizado no Oak Ridge National Laboratory.

Sequoia - IBM

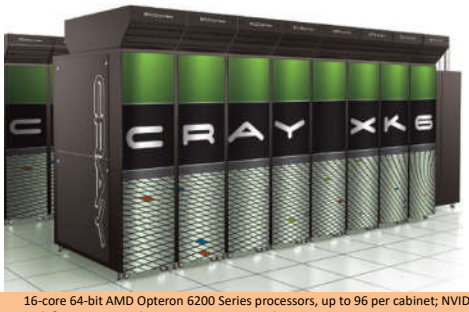
Possui 1,6 milhão de núcleos de processamento e 1,6 petabyte (ou pouco mais de 1,6 mil terabytes) de memória RAM, sendo resfriado em sua maior parte por água. Executa 16,32 quadrilhões de cálculos por segundo, a IBM deu um exemplo bem interessante: três bilhões de pessoas usando uma calculadora de bolso precisarão realizar um milhão de operações por segundo para atingir um processamento equivalente ao Sequoia.

Sequoia

Blue Gene – IBM

It was first delivered to the Lawrence Livermore National Laboratory in 2011and now full deployed with an impressive 16.32 Petaflop/s on the Linpack benchmark using 1,572,864 cores.

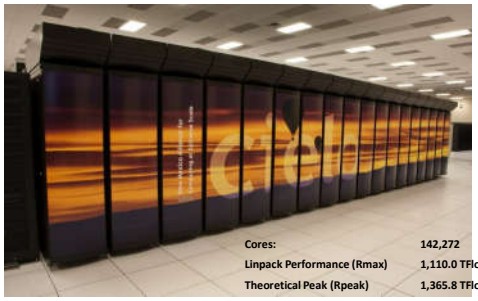
CRAY – XK6



Processor	16-core 64-bit AMD Opteron 6200 Series processors, up to 96 per cabinet; NVIDIA® Tesla® K20 GPU Accelerators, up to 96 per cabinet
Memory	16 GB or 32 GB registered ECC DDR3 SDRAM and 6 GB GDDR5 per compute node Memory bandwidth: 4 channels of DDR3 memory per compute node
Compute Cabinet	AMD processing cores: 1,536 processor cores per system cabinet Peak performance: 100+ Tflops per system cabinet

Cielo

Cielo is a supercomputer located at Los Alamos National Laboratory in New Mexico, USA built by Cray Inc.



Cores:	**142,272**
Linpack Performance (Rmax)	**1,110.0 TFlop/s**
Theoretical Peak (Rpeak)	**1,365.8 TFlop/s**
Power:	**3,980.00 kW**

Jaguar - CRAY

o Jaguar, do Laboratório Nacional de Oak Ridge, no Departamento de Energia dos Estados Unidos.Esta máquina alcança 1,75 petaflop/s.
O Jaguar tem 224.162 núcleos e usa sistema operacional Linux, tendo sido instalado em 2009 no Laboratório Nacional de Oak Ridge, do Departamento de Energia dos Estados Unidos.