CAP-387(2016) – Tópicos Especiais em Computação Aplicada: Construção de Aplicações Massivamente Paralelas

Aula 3: Sistemas Massivamente Paralelos Atuais - Brasil

Celso L. Mendes, Stephan Stephany

LAC /INPE

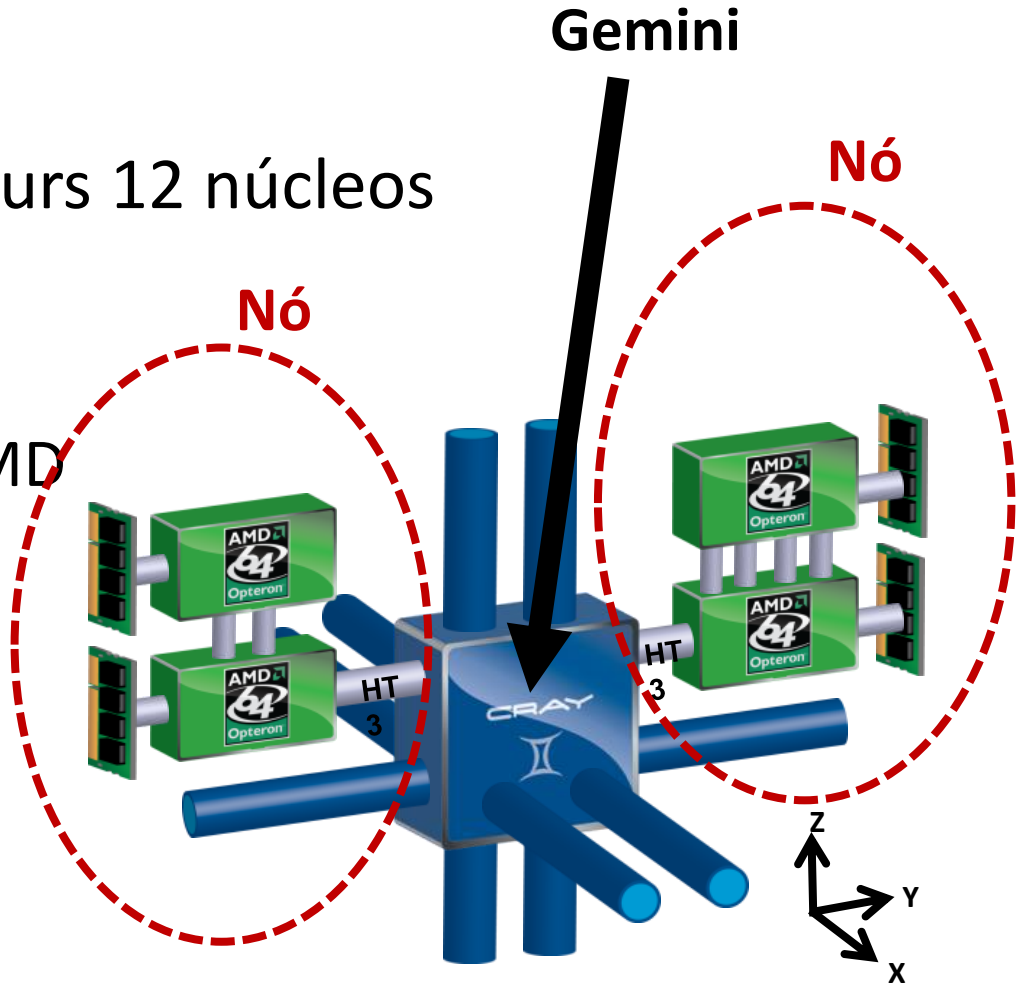
Emails: celso.mendes@inpe.br, stephan.stephany@inpe.br

Sistema Tupã – CPTEC/INPE

- Cray XE6 - Previsão de tempo e clima
  - Instalação: 2010 (Cachoeira Paulista)
  - Desempenho de pico ≈ 258 Tflops/s
  - #29 na lista Top500 de Novembro/2010
  - 14 racks, 96 nós em cada rack
  - ~1.300 nós de computação, 2 AMD-Opteron por nó
    - 24 núcleos em cada nó, mais de 30.000 núcleos total
    - Adicionalmente, 40 nós de serviço
  - Rede de interconexão: Gemini (torus-3D)
    - Desenvolvida pela Cray, baseada em FPGAs

Sistema Tupã – CPTEC/INPE

- Cray XE6:
  - Nó: 2 x CPU Opteron Magny-cours 12 núcleos
  - 2 nós por Gemini
  - Memória: 32 Gbytes por nó
    - 16 Gbytes por processador AMD
    - Esquema NUMA
  - Não há disco local
    - Discos são externos ao torus
    - I/O feito via nós de serviço
    - I/O também utiliza a Gemini



Sistema Tupã – CPTEC/INPE

- Outros atributos
  - Racks iguais ao…
    - Blue Waters (288)
    - Titan (200)
  - Placas similares (só CPUs)
  - Saída da Top500: Jun/2016
  - Em final de vida útil
    - ~6 anos de bons serviços!
    - XE/XK fora de fabricação

Sistema Santos Dumont - LNCC

- Principais Características
  - Desempenho agregado de pico acima de 1 Pflops/s
  - Fabricante: Bull (França)
  - Instalação física: containers

Sistema Santos Dumont (cont.)

- Arquitetura:
  - Diversidade de processadores: CPU, GPU, Xeon-Phi
  - 3 sub-sistemas listados no Top-500 de Junho/2016:
    - #265 (GPU): 658 TF, #364 (Phi): 479 TF, #433 (CPU): 348 TF
  - Quatro tipos de nós computacionais:
    a) 504 nós com 2 CPUs Intel Xeon – 24 núcleos/nó
    b) 198 nós com 2 CPUs Intel Xeon e **2 GPUs Nvidia K40**
    c) 54 nós com 2 CPUs Intel Xeon e **2 Intel Xeon-Phi (KNC)**
    d) 1 nó com **16** CPUs Intel Xeon (240 núcleos), 6 TB de RAM
  - Rede de interconexão: Infiniband

Sistema Santos Dumont (cont.)

- Ambiente de Software:
  - Configurado através de “módulos”
    - *module load/unload/list/avail/show/whatis*
  - Compiladores: GNU, Intel, PGI, Nvidia
  - MPI: OpenMPI, Intel
  - Sistema Operacional: Linux
  - URL: http://sdumont.lncc.br/support_manual.php?pg=support#
  - Suporte: enviar email para *helpdesk-sdumont@lncc.br*

Sistema Santos Dumont (cont.)

- Submissão e Controle de Jobs:
  - Gerenciador de filas e tarefas: Slurm v.14.11
  - Filas mais relevantes para este cur*so:*
    - *cpu_dev* - limites: até 2 horas, até 20 nós
    - *cpu* - limites: até 48 horas, até 50 nós (20 na prática)
  - Outras filas de possível interesse: GPU e Xeon-Phi
    - *nvidia_dev* – limites: até 2 horas, até 2 nós
    - *nvidia* – limites: até 48 horas, até 50 nós (20 na prática)
    - *phi_dev* – limites: até 2 horas, até 2 nós
    - *phi* – limites: até 48 horas, até 50 nós (20 na prática)

Sistema Santos Dumont (cont.)

- Submissão e Controle de Jobs (cont.):
  - Submissão de jobs: através de scripts
  - Exemplos em http://sdumont.lncc.br/support_manual.php?pg=support#6
  - Comando de submissão: *sbatch <script>*
  - Verificação de status*:* *squeue*
  - Cancelamento de jobs: *scancel <jobid>*
  - Execução interativa: comando *salloc*
    - Permite éxecuções manuais, ao invés de via script
    - Ver manual acima, para opções e formato
    - Uso deve ser cuidadoso, para minimizer gastos do projeto

S.Dumont – Exemplo de Script

```
#!/bin/bash
#SBATCH --nodes=10
#SBATCH --ntasks-per-node=24
#SBATCH --ntasks=240      #Numero total de tarefas MPI
#SBATCH -p cpu          #Fila (partition) a ser utilizada
#SBATCH -J meujob         #Nome do job
#SBATCH --time=00:02:00   #Tempo limite
#SBATCH --exclusive       #Utilizacao exclusiva dos nós
#Exibe os nós alocados para o Job
echo $SLURM_JOB_NODELIST
nodeset -e $SLURM_JOB_NODELIST
cd $SLURM_SUBMIT_DIR
#Configura I_MPI_PMI_LIBRARY para apontar para a biblioteca "Process Management Interface" do Slurm
export I_MPI_PMI_LIBRARY=/usr/lib64/libpmi.so
#Define o executavel e dispara execução
EXEC=/scratch/charm/celso.mendes/pname
srun -n $SLURM_NTASKS $EXEC
```