NCCS Cluster Architecture
Series of Brown Bag Presentations
October 2006

NASA Center for Computational Sciences (NCCS)
Computational & Information Sciences and Technology Office (CISTO)
Goddard Space Flight Center

So we’re going to be roommates?

- I’m a 4-year old cluster
- I have 100 cabinets
- I have 384 nodes
- I have 24 TB of storage in 12 racks
- I have over 11 miles of cables and 10 switches
- I require 100 tons of cooling
- Yeah, but my peak computing capacity is 3.2 TF

Picture courtesy of the Apple web site. No permission was given by Apple to use this picture.

- I’m a brand new cluster
- I have only 5 cabinets
- I have 128 nodes
- I have 60 TB of storage in one-half of a rack
- I have only 0.5 miles of cables and 2 switches
- I require about 30 tons of cooling
- Well, my peak computing capacity is 3.3 TF

How much heat?

Common Appliances		Heat ~BTU/hr	Old Cluster ~1,200,000 BTU/Hr	New Cluster ~260,000 BTU/Hr
Common Toaster		5,000	240	52
Window Air Conditioner		10,000	120	26
Wood Stove		35,000	34.3	7.4
Frymaster MJ35-SDN Deep Fryer 65 lbs of French fries per hour		110,000	10.9 709 lbs of frozen French fries per hour	2.4 154 lbs of frozen French fries per hour

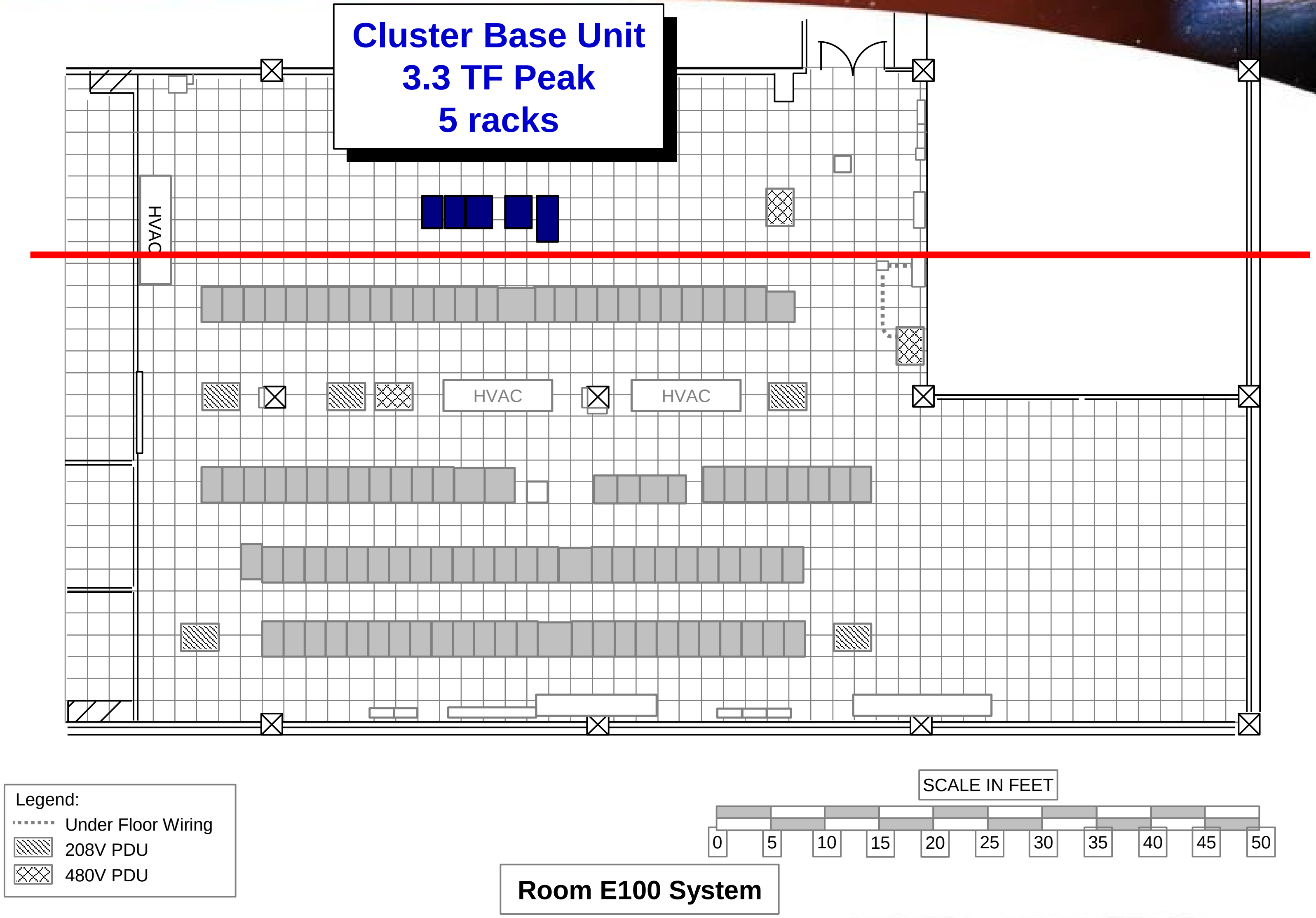
It’s not the total heat that kills you...

- Old Cluster
  - ~1,200,000 BTU/Hr
  - 100 racks = 600 sq ft
  - **2,000 BTU/Hr/sq ft**
  - ~5 KW maximum rack power

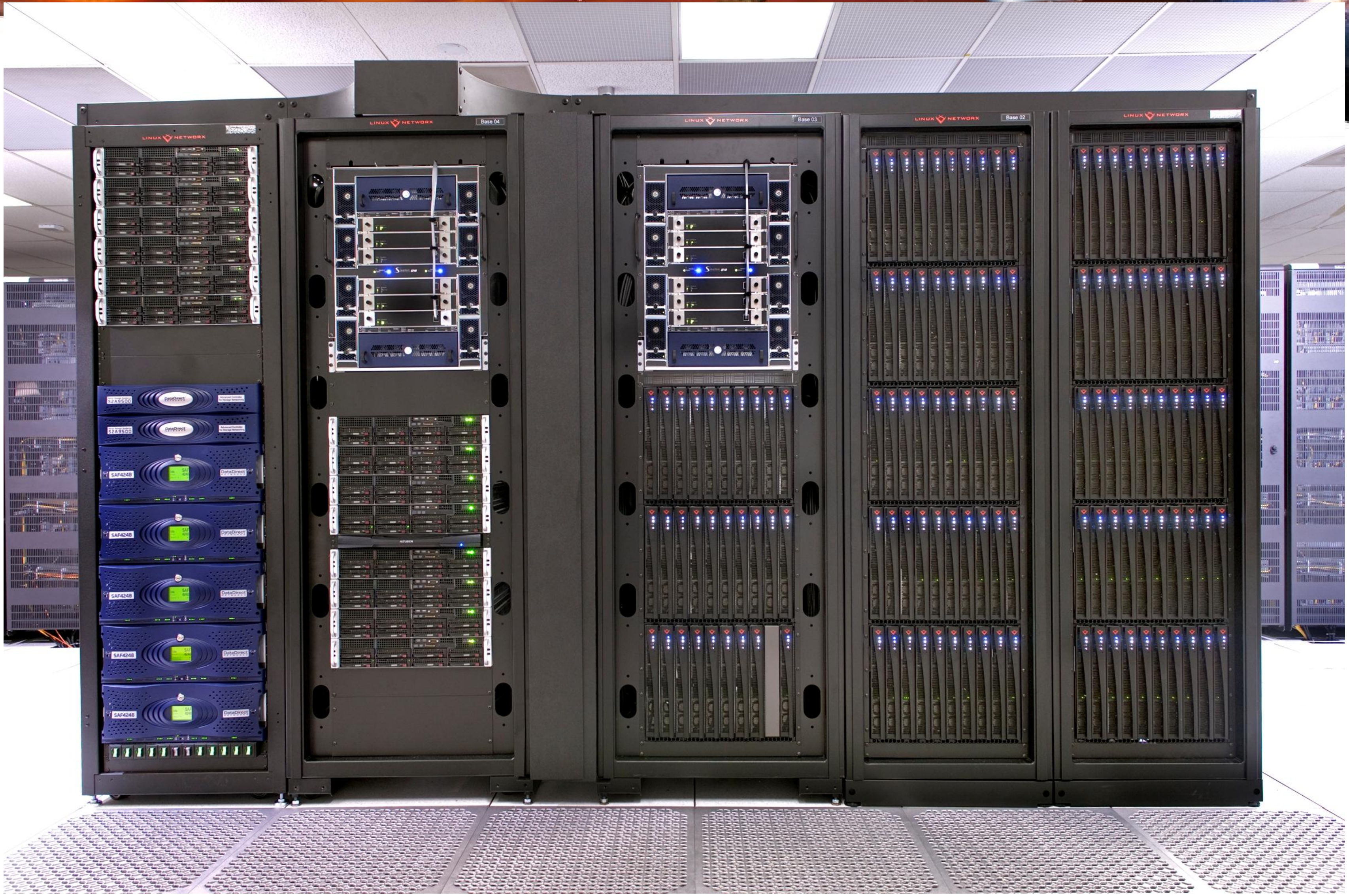
- New Cluster
  - ~260,000 BTU/Hr
  - 5 racks = 30 sq ft
  - **~8,600 BTU/Hr/sq ft**
  - 25 KW maximum rack power draw

Increase of over 4 to 5x of the heat per unit area and maximum power per rack.

Roommates – Don’t cross this line!



System is actually here!



Partners in Crime

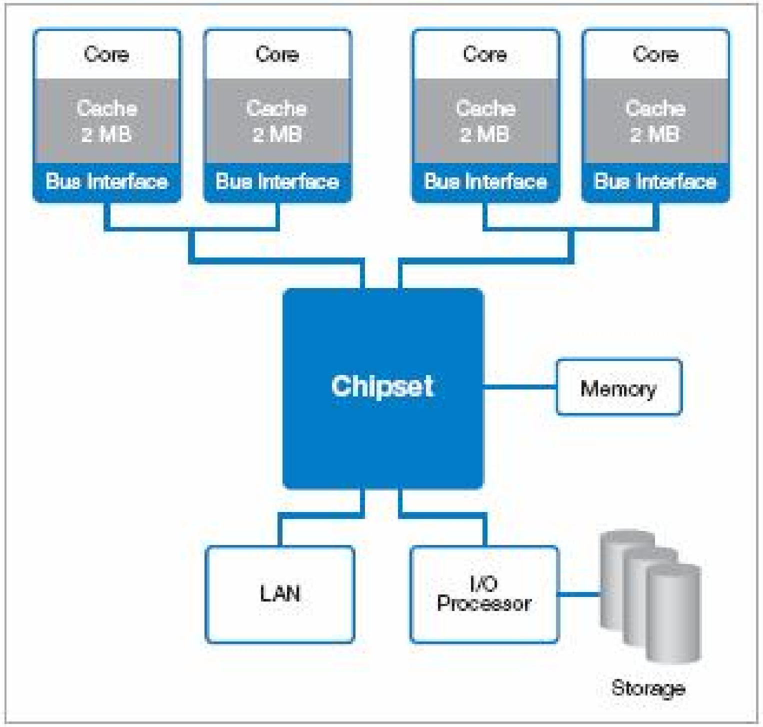
- Linux Networx
  - ww.lnxi.com
- Intel
  - www.intel.com
- SilverStorm
  - www.silverstorm.com
- Data Direct Networks
  - www.datadirectnet.com
- IBM
  - www.ibm.com
- Altiar (PBS)
  - www.altair.com
- Computer Sciences Corporation
  - www.csc.com

Compute Node Architecture

- Base Unit
  - SuperMicro mother board, 0.8U Evolocity II chassis
  - Dual socket, dual core Intel Dempsey 3.2 GHz
  - 120 GB hard drive
  - 4 GB RAM (4 x 1 GB DDR2 533 MHz FB DIMM)
  - PCI-Express with SilverStorm inifinband 4x HCA (10 Gb)
- Scalable Unit
  - Dell mother board, 1U chassis
  - Dual socket, dual core Intel Woodcrest 2.66 GHz
  - 160 GB hard drive
  - 4 GB RAM (4 x 1 GB DDR2 667 MHz FB DIMM)
  - PCI-Express with SilverStorm inifinband 4x HCA (10 Gb)

Dempsey Architecture

- Intel Dempsey 3.2 GHz
  - Dual socket, dual core (4 cores per node)
  - 2 x 64-bit floating point operations per clock cycle (per core)
  - 2 MB L2 cache per core (4 MB total per socket)
  - Cache speed? Cache line size?
  - 4 GB/s memory bandwidth to the core (peak)
- Peak Computing
  - 12.8 GF per socket, 25.6 GF per node
  - Compare that to the 6 GF per socket for the Itanium processors on Explore
  - Early indications with High Performance Linpack (HPL) are showing 60% - 65% of peak



Infiniband Interconnect Architecture

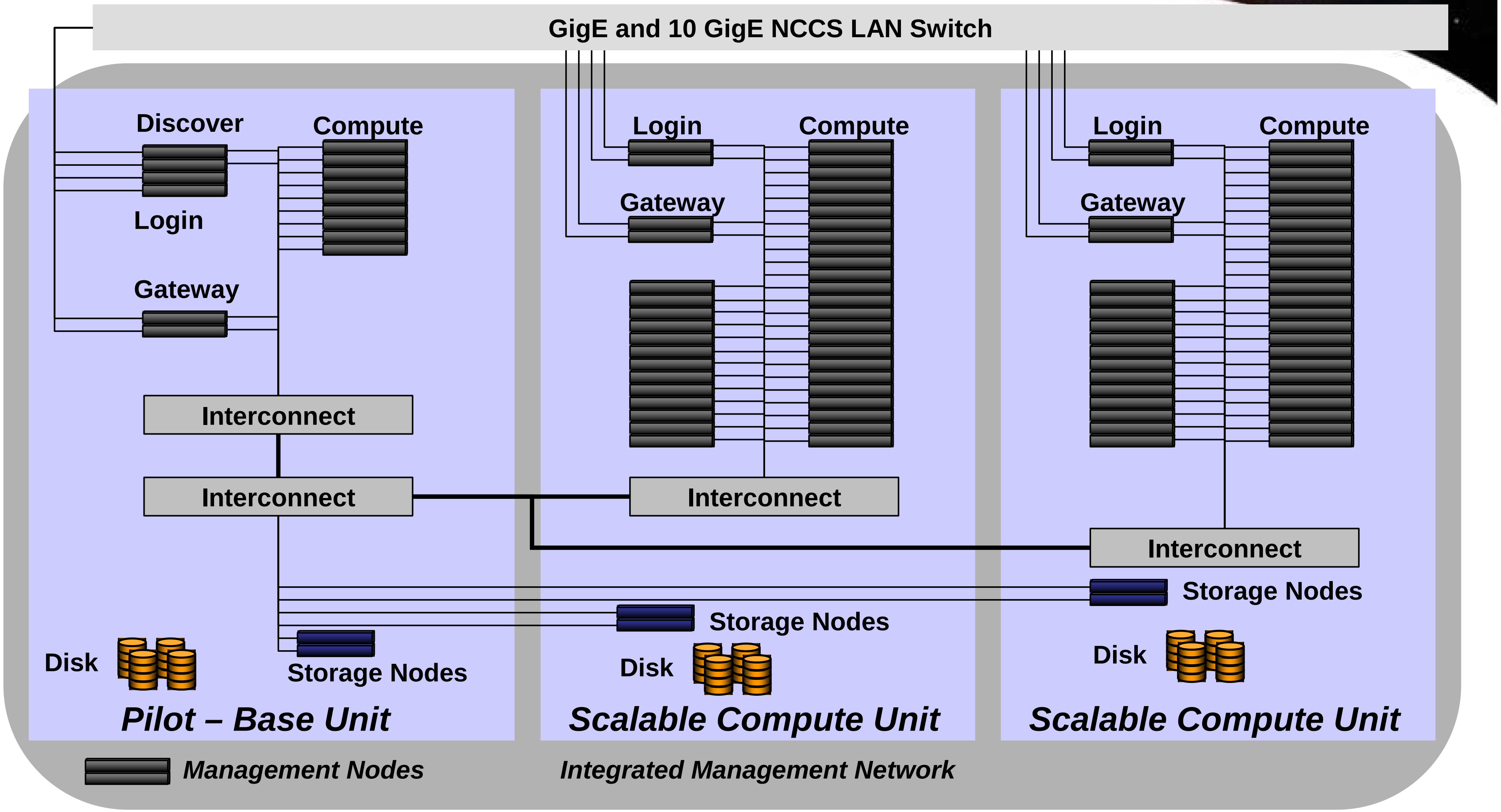
- SilverStorm Infiniband Switches
  - 9240 switch chassis
  - Up to 288 ports per chassis
  - SilverStorm IB software stack, moving to the Open Fabrics software stack in the future
- Mellanox InfiniHost III Ex dual-ported 4x Infiniband Host Channel Adapters (HCA)
  - PCI-Express 8x
  - Double data rate
  - 20 Gb/s bi-directional
- Of interest...
  - SilverStorm (www.silverstorm.com) was recently acquired by QLogic

Cluster File System

- IBM Global Parallel File System (GPFS)
  - All systems run a client
  - Clients cache metadata (causes some memory overhead)
  - Separate data servers (NSD) and metadata servers (MDS)
- Data Direct Network (DDN) SATA for Data
  - 500 GB 7200 RPM drives
  - 60 TB raw for base unit
  - Will increase by approximately 90 TB raw with the addition of EACH scalable unit
- Engenio FC for Metadata
  - 146 GB 15K RPM drives
  - Highly redundant

Discover…the Borg

“You will be assimilated”



Logging in...

- Log into discover
  - ssh to login.nccs.nasa.gov
  - Enter SecurID pin number and code
  - Choose discover as your host
  - Enter password
  - DNS will round robin users between the four (4) discover nodes
    - discover0[1-4]
- Compute nodes will use the hostnames “borg”
  - Base unit compute nodes will follow the following convention: borga###
  - Hence, the compute nodes will be borga001 through borga130
  - As additional scalable units come into the cluster, the compute nodes will be designated with borgb###, borgc###, etc.

Software Stack

- SUSE Linux
  - 9 service pack 3
  - Moving to 10 sometime in the future (probably about the same time as the Altix systems)
  - Must have GPFS support prior to upgrades
- Compilers
  - Intel, PGI, gcc
  - Coming later: PathScale, Absoft
- MPI
  - Intel, Scali, SilverStorm
- OpenMP
  - Intel
- Tools
  - Totalview, Intel vtune, Intel trace analyzer