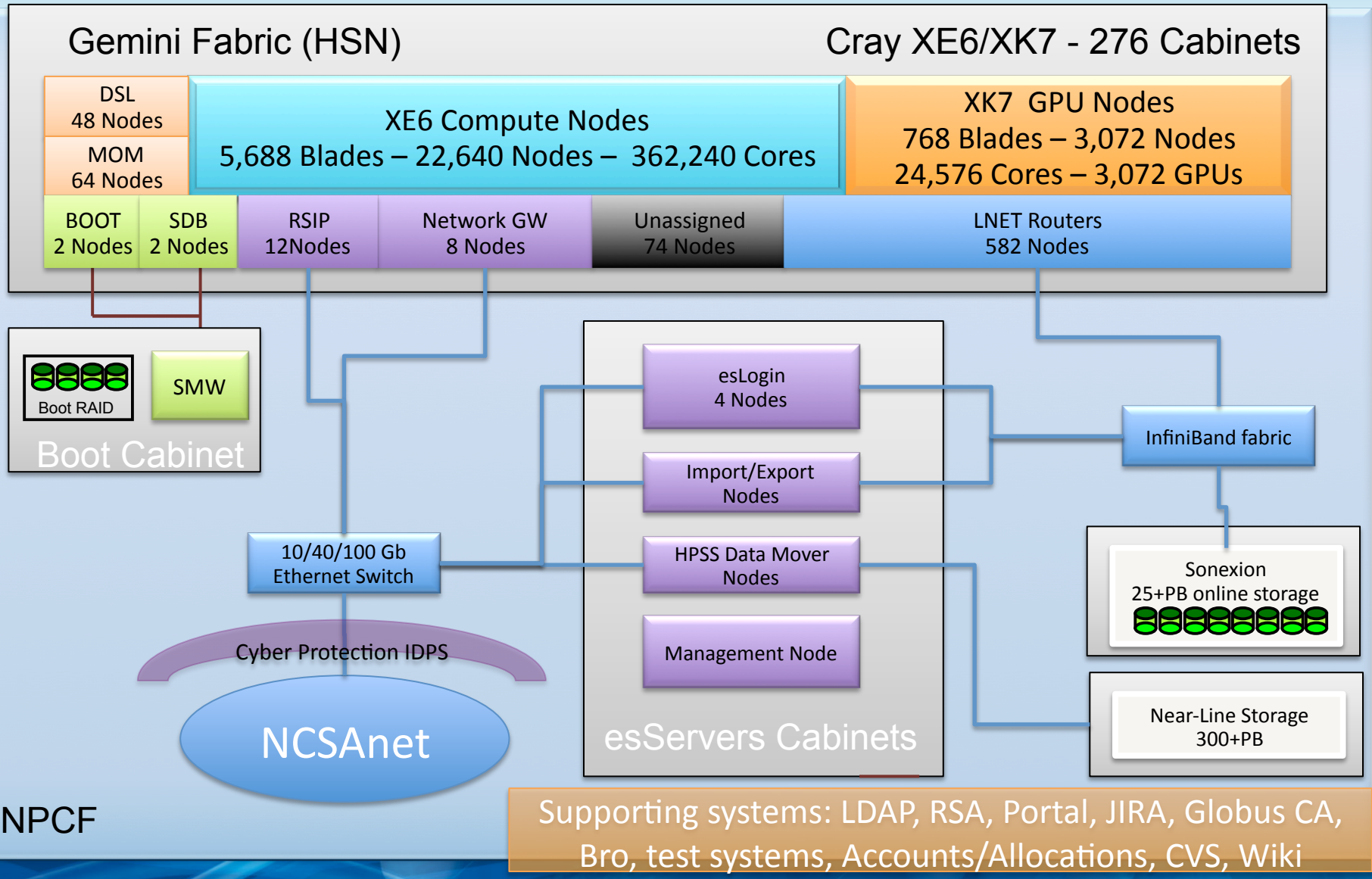
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Algorithm and Software Needs at Extreme Scale

William Gropp

GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION


Gemini Fabric (HSN)
Cray XE6/XK7 - 276 Cabinets
DSL
48 Nodes
MOM
64 Nodes
XE6 Compute Nodes
5,688 Blades – 22,640 Nodes – 362,240 Cores
XK7 GPU Nodes
768 Blades – 3,072 Nodes
24,576 Cores – 3,072 GPUs
BOOT
2 Nodes
SDB
2 Nodes
RSIP
12Nodes
Network GW
8 Nodes
Unassigned
74 Nodes
LNET Routers
582 Nodes
Boot RAID
SMW
Boot Cabinet
esLogin
4 Nodes
Import/Export
Nodes
HPSS Data Mover
Nodes
Management Node
esServers Cabinets
InfiniBand fabric
10/40/100 Gb
Ethernet Switch
Cyber Protection IDPS
NCSAnet
Sonexion
25+PB online storage
Near-Line Storage
300+PB
NPCF
Supporting systems: LDAP, RSA, Portal, JIRA, Globus CA,
Bro, test systems, Accounts/Allocations, CVS, Wiki

System Attribute	NCSA Blue Waters	LLNL Sequoia (#1)
Vendor(s)	Cray/AMD/NVIDIA	IBM
Processors	Interlagos/Kepler	PowerPC A2
Total Peak Performance (PF)	11.9	20.1
Total Peak Performance (CPU/GPU)	7.6/4.3	20.1/0.0
Number of CPU Chips (8, 16 cores/chip)	48,576	98,304
Number of GPU Chips	3,072	0
Amount of CPU Memory (TB)	1,510	1,572
Interconnect	3-D Torus	5-D Torus
Amount of On-line Disk Storage (PB)	26	50(?)
Sustained Disk Transfer (TB/sec)	>1	0.5-1.0
Amount of Archival Storage	300	?
Sustained Tape Transfer (GB/sec)	100	?

Scaling and Performance Issues

- We believe that future systems will have many features in common with Blue Waters:
  - Heterogeneous processing elements
    - In BW, these are NVIDIA GPUs and AMD CPUs
    - In future systems, expect these to be more tightly integrated
    - Can also view vector instructions and AMD core/ core module as a heterogeneous architecture
  - Need algorithms and software that can fully exploit complex processing elements

Scaling and Performance Issues

- Multilevel communication topology
  - While flatter communication networks are planned/ likely in the next generation of systems, issues of intra/inter node remain
  - Avoiding hotspots and communication contention important now and in the future, though details will depend on network topology
  - Need
    - Algorithms that can adapt to topology
    - System software to discover/exploit it

Scaling and Performance Issues

- Performance irregularity in SMPs requires more adaptive algorithms and software
  - Need to restructure algorithms to be more adaptive, not assume performance regularity
  - Need software and programming support to make it easier to
    - Program
    - Diagnose/repair performance issues

Data and Big Data

- Definition of Big Data: Does not fit into memory
  - For us, that is > 1.5PB
- Looking for applications that can *and must* use this (nearly) unique capability
  - Not typical MapReduce problems
  - E.g., need random access to entire data set, not stream access to organized subsets.
- Once we have some compelling applications, usual issues: Algorithms and Programming

Illinois Vision of Collaboration

- Bring Complementary Strengths to bear on problems of extreme scale
- Exploit availability of Blue Waters (now in acceptance test phase)
  - More details on Blue Waters tomorrow…