# BLUE WATERS

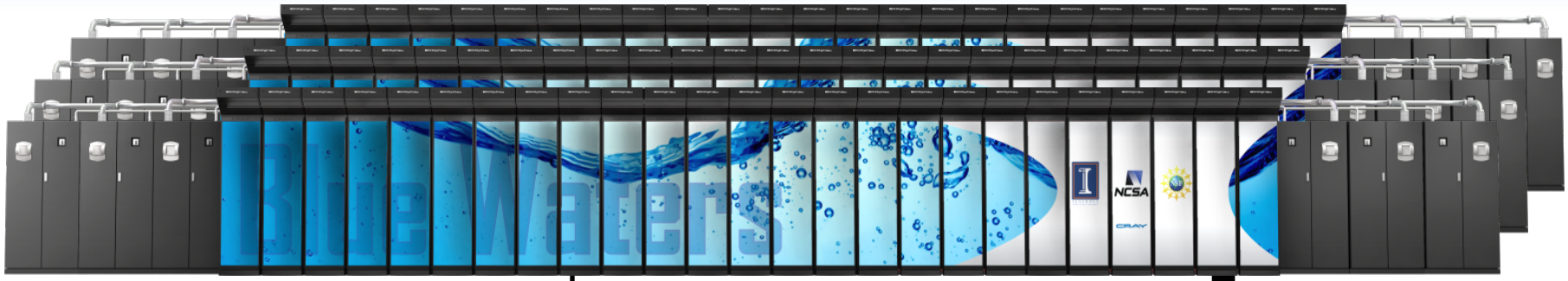
SUSTAINED PETASCALE COMPUTING

## Is Petascale Complete? What Do We Do Now?

Dr. William Kramer
National Center for Supercomputing Applications, University of Illinois


NCSA NSF GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

# Blue Waters Computing System




Aggregate Memory – 1.6 PB

10/40/100 Gb Ethernet Switch

External Servers

IB Switch

>1 TB/sec

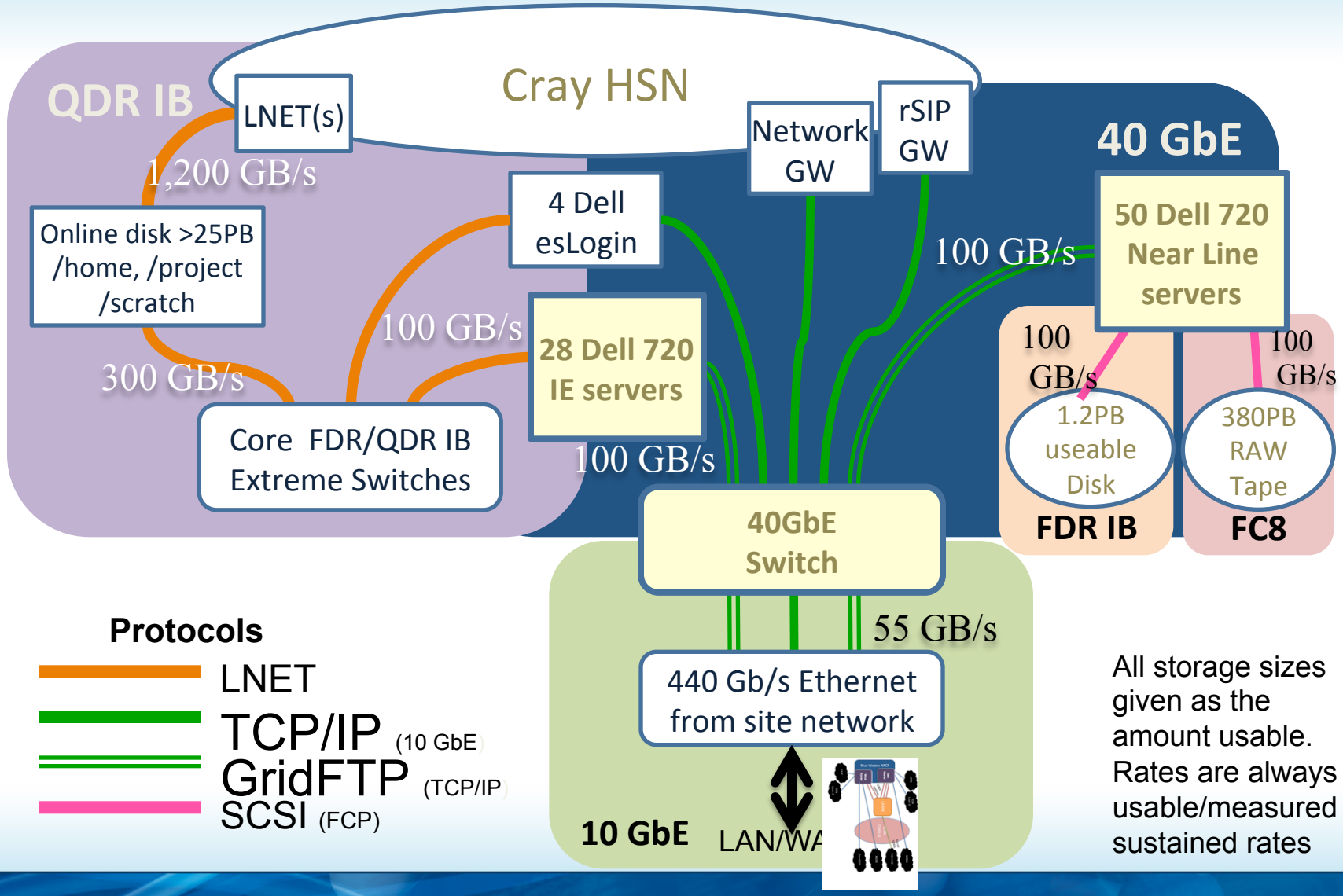
120+ Gb/sec

100 GB/sec

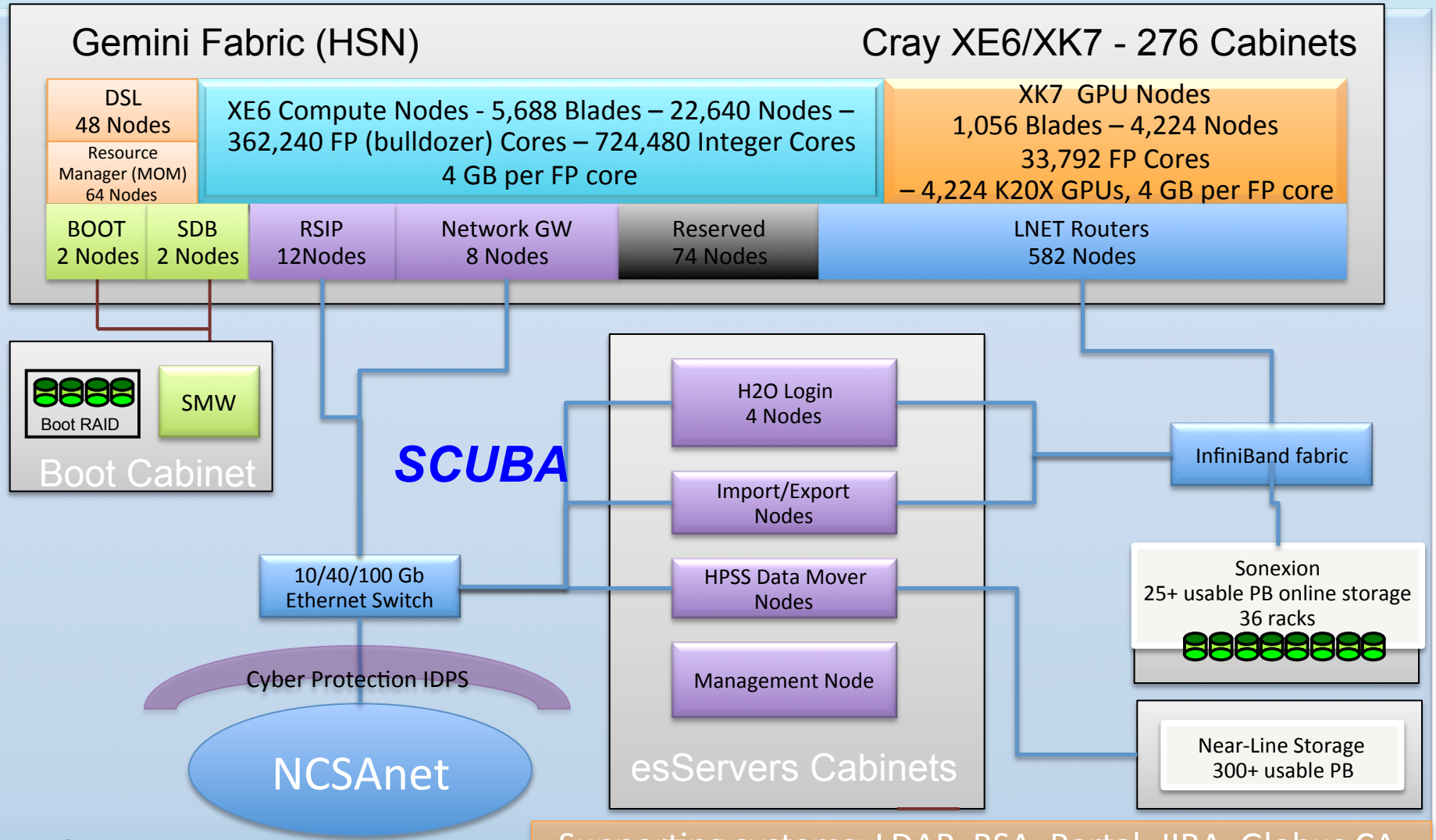
100-300 Gbps WAN

Spectra Logic: 300 usable PB

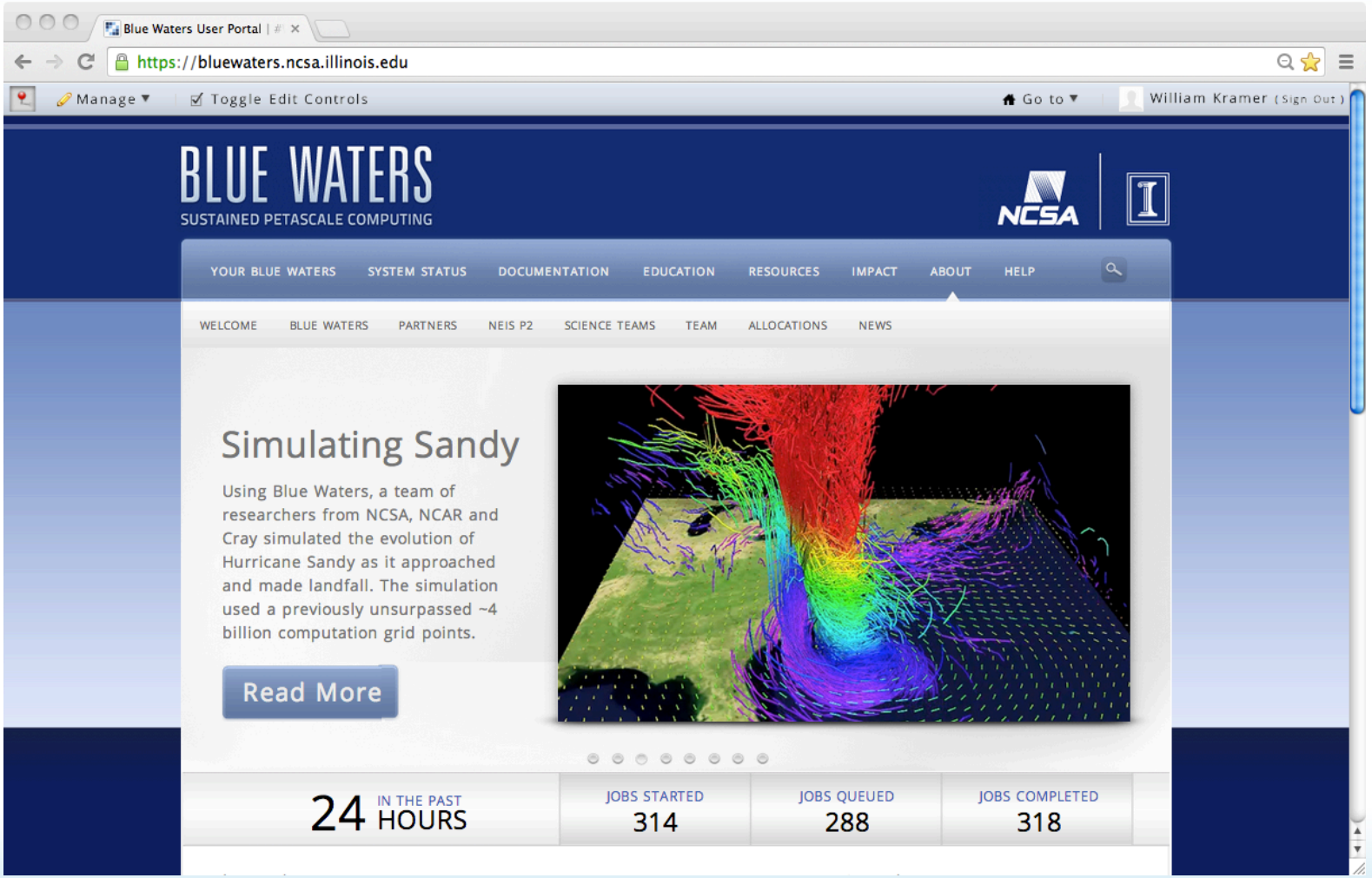
Sonexion: 26 usable PB



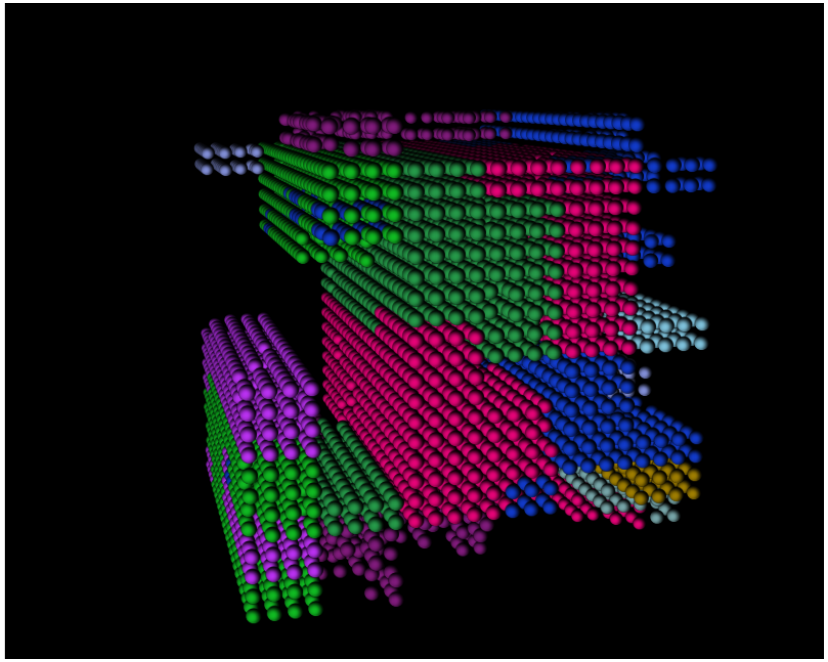

All storage sizes given as the amount usable. Rates are always usable/measured sustained rates

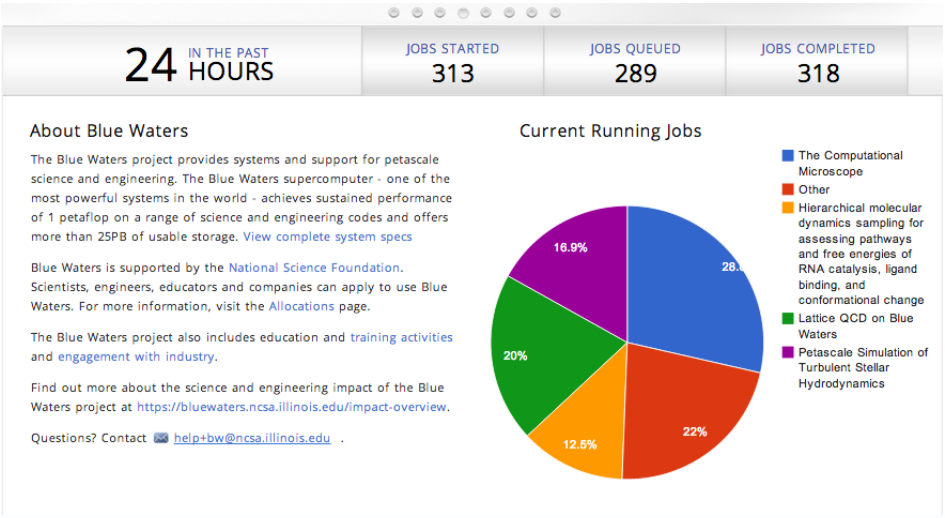

Gemini Fabric (HSN)
Cray XE6/XK7 - 276 Cabinets
DSL
48 Nodes
Resource Manager (MOM) 64 Nodes
XE6 Compute Nodes - 5,688 Blades – 22,640 Nodes – 362,240 FP (bulldozer) Cores – 724,480 Integer Cores
4 GB per FP core
XK7 GPU Nodes
1,056 Blades – 4,224 Nodes
33,792 FP Cores
– 4,224 K20X GPUs, 4 GB per FP core
BOOT
2 Nodes
SDB
2 Nodes
RSIP
12Nodes
Network GW
8 Nodes
Reserved
74 Nodes
LNET Routers
582 Nodes
Boot RAID
SMW
Boot Cabinet
SCUBA
H2O Login
4 Nodes
Import/Export Nodes
HPSS Data Mover Nodes
Management Node
esServers Cabinets
InfiniBand fabric
10/40/100 Gb Ethernet Switch
Sonexion
25+ usable PB online storage
36 racks
Cyber Protection IDPS
NCSAnet
Near-Line Storage
300+ usable PB
NPCF
Supporting systems: LDAP, RSA, Portal, JIRA, Globus CA, Bro, test systems, Accounts/Allocations, CVS, Wiki

# http://bluewaters.ncsa.illinois.edu



# http://bluewaters.ncsa.illinois.edu





# IS PETASCALE COMPLETE?

Of course not.

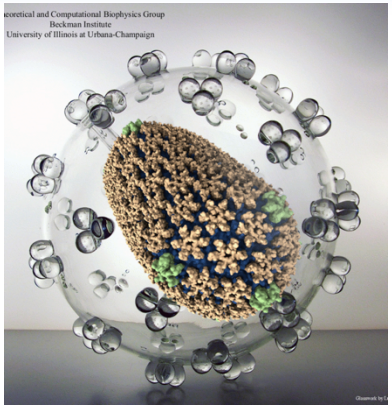
Petascale is not complete until much Petascale Science and Engineering is successful.

But we can start to draw conclusions

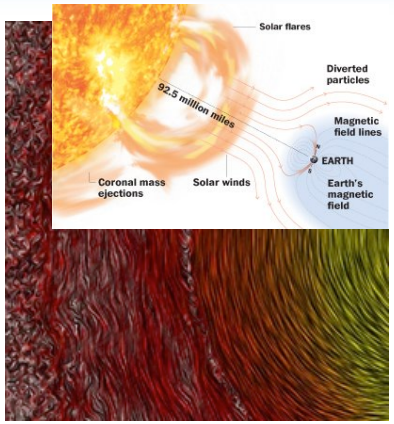
- See other talks for tomorrow

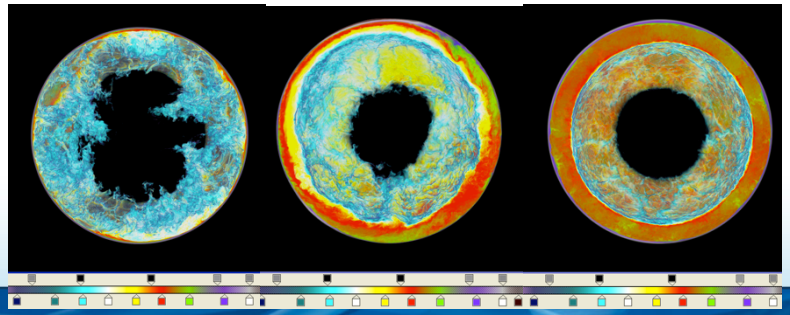
# PETASCALE LESSONS THAT @SCALE MUST ADDRESS

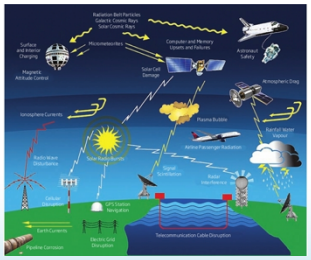
# 1 – Petascale Works




| Category | Number of Teams |
|---|---|
| NSF - PRAC | 28 active<br>+6 exploratory<br>5 have completed |
| University of Illinois | 30<br>15 General, 15 Exploratory |
| GLCPC | 10 |
| Education | 4 |
| Industry | 1 |
| Innovation and Exploration | 8 |





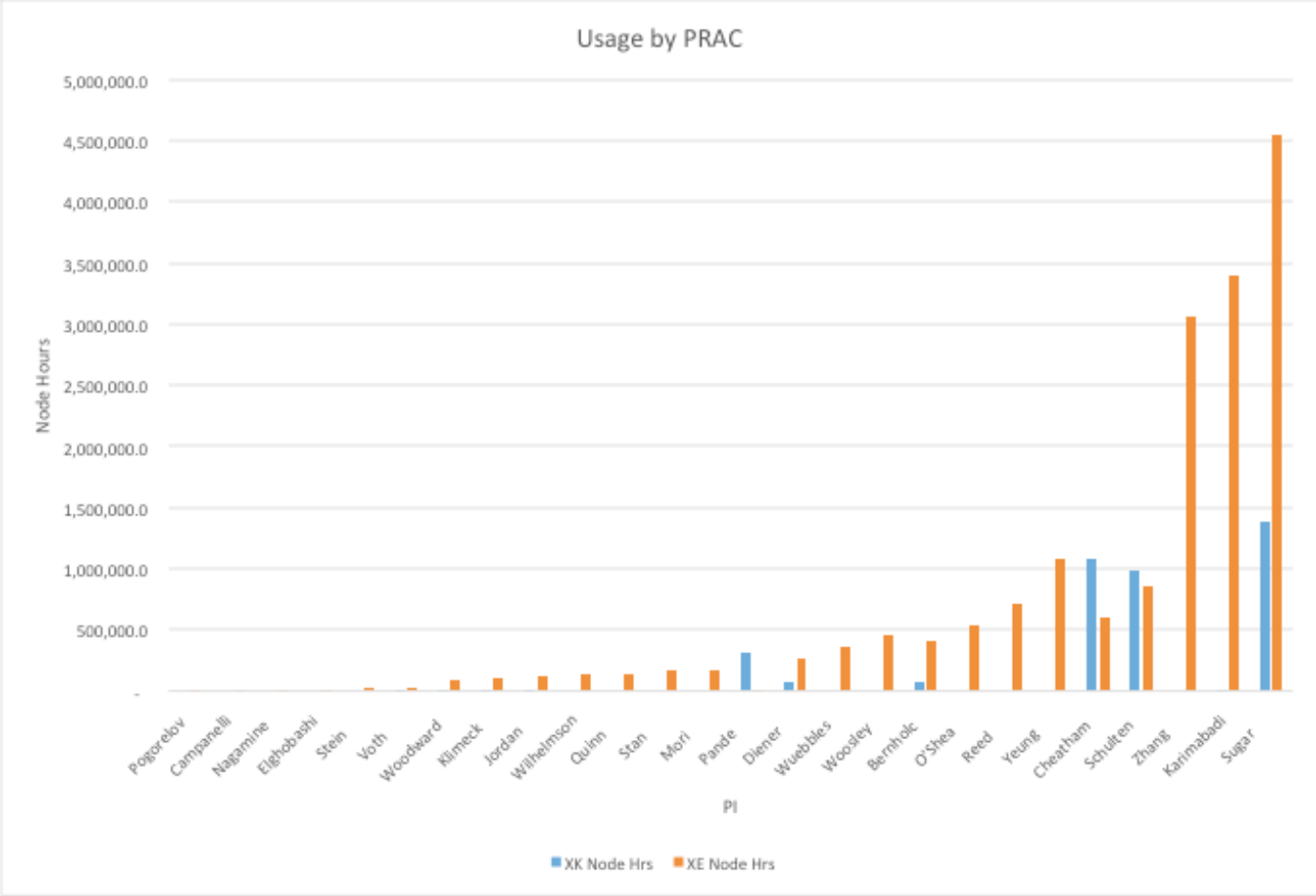
# 1 – It Works

- Petascale Definitions of Scale
  - Not Large ≤ 1,284 nodes
    - ≤ 20,544 FP cores
    - ≤ 41,088 integer cores
  - Large ≥ 1,285 nodes
    - ≥ 20,560 FP cores
    - ≥ 41,120 integer cores
  - Very Large ≥ 4,584 nodes
    - ≥ 123,344 FPcores
    - ≥ 146,688 integer cores
- Year to Date Computational Usage
  - Not Large - 60%
  - Large - 25%
  - Very Large - 15%
    - Does not include any GPU usage

No longer can define core, processor…
- ~380,000 AMD x86 Floating-point Bulldozer cores,
- ~760,000 AMD x86 integer cores,
- 4,224 NVIDIA Kepler K20x GPUs or
- >12 million “cuda-cores”

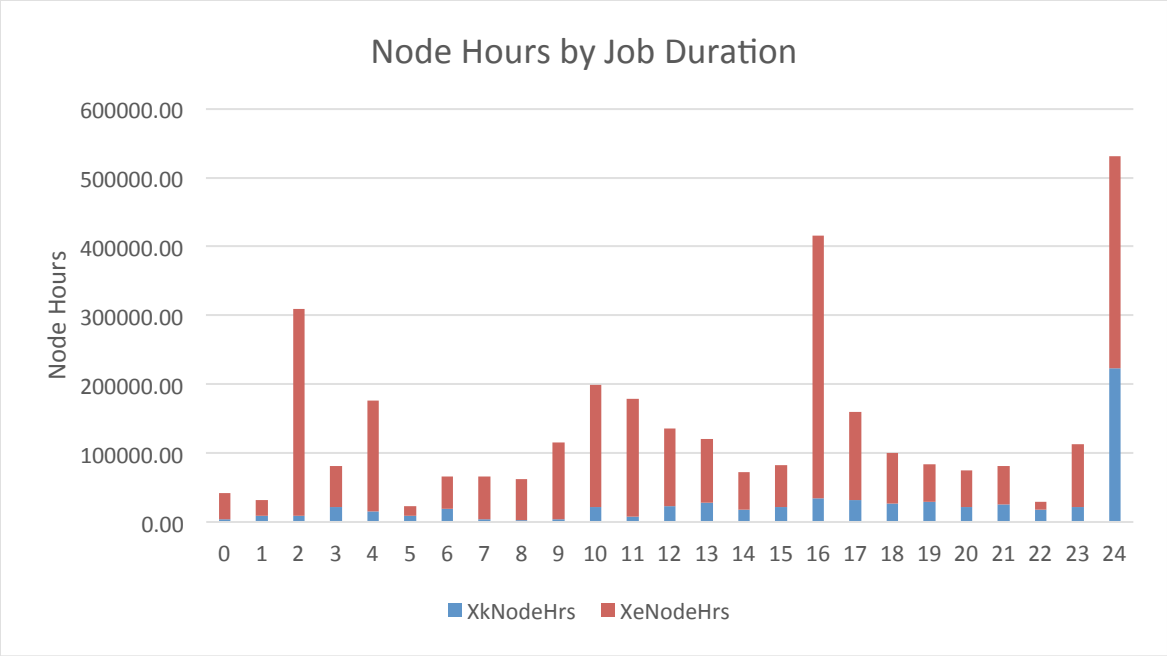
# 1 – It Works - Usage by NSF PRAC team

An observed experiment – teams self select what is most useful



# 1 – It Works Interim Full Service Job Characteristics

- July-Sept. Interval
- Large job expansion factor well under target of 10.
- 1+(time in queue/ time requested)




| | Small | Medium | Large |
|---|---|---|---|
| XE nodes | 1- 1,132 nodes | 1,133 - 4,528 nodes | 4,529 - 25,712 nodes |
| XK nodes | 1 - 16 nodes | 17 - 256 nodes | 257 – 4,224 nodes |

| Expansion factor | Large jobs | Medium jobs | Small jobs |
|---|---|---|---|
| XK nodes | 1.56 | 2.85 | 2.79 |
| XE nodes | 4.75 | 1.27 | 1.04 |

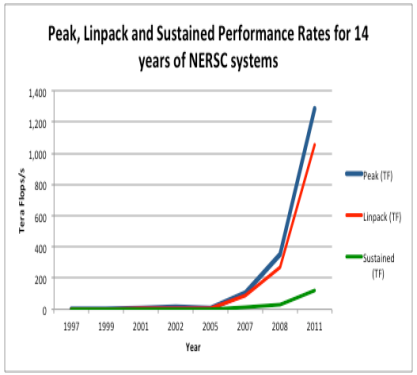
# 2 – It Takes Long Term Efforts

- 2006 – Planning, community workshops
- 2007 – Call for Proposal, Response, Selection
- 2008 – Award, Contracts, Project Formalization
- 2009 – Construction, Co-Design with Application Teams, Early Systems, Software Development
- 2010 – NPCF Completes, Co-Design with Application Teams, Networking Infrastructure Installation, Software Development
- 2011 – Technology “Grand Pivot”, Delivery of First Cray Components
- 2012 – Full Cray System Installation, Near-line Storage Installation, Testing, Early Use, Acceptance
- 2013 – Early Production, Upgrade, Full Production
- 2013-2018 – Production Science

# 3 – Sustained Performance At Scale Achievable NEIS-P$^2$ – Direct Support

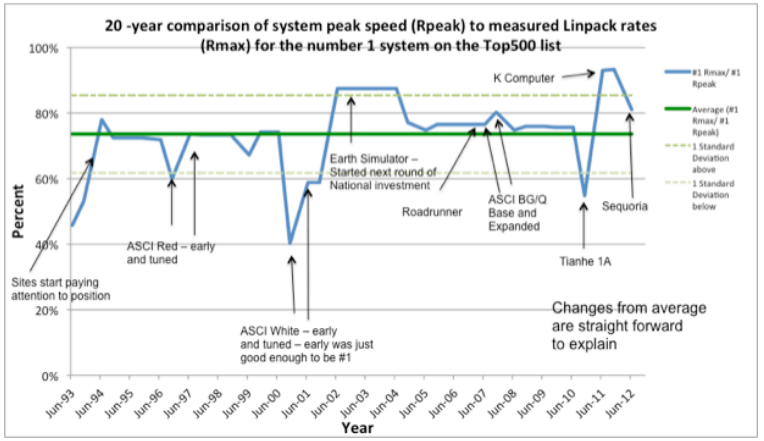
- NEIS-P$^2$ – Direct Support
  - Blue Waters directly funded science teams to make improvements in their codes to enable them to "realize the full potential of the Cray XE6/XK7 system."
  - 20 PRAC teams participated
  - Component was **completed in Summer 2013**
  - Summary of activities and results for each team can be found on the Blue Waters website: https://bluewaters.ncsa.illinois.edu/neis-p2-final-reports
- SPP Measure for Blue Waters – 1.31 PF/s
  - 12 full application time to solution tests
  - 6 applications ran on entire system above 1 PF/s during acceptance
    - Full time to solution
- Reporting by full Science Teams indicated more applications

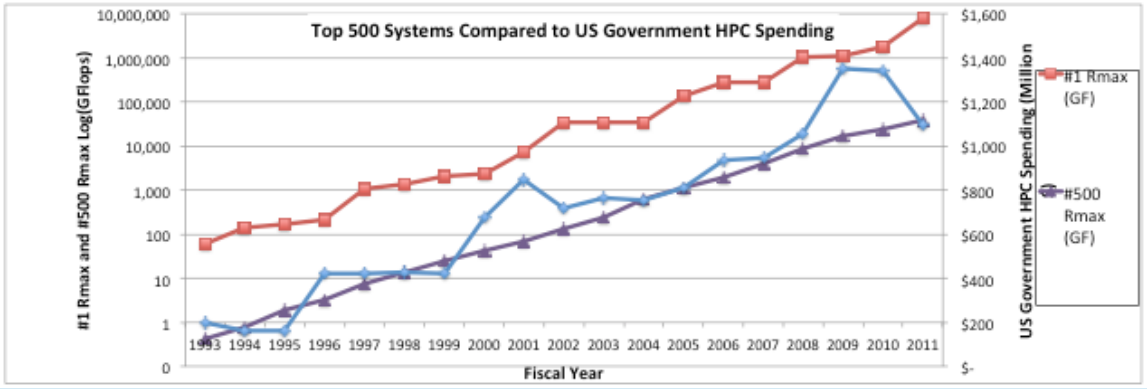
# 4 – There is Life Beyond the Top500




Top500 values do not correlate with vs measured System Sustained Performance - 13 years of systems at NERSC show this trend

TOP500 is dominated by who has the most money to spend–not what system is the best.



# 4 – There is Life Beyond the Top500

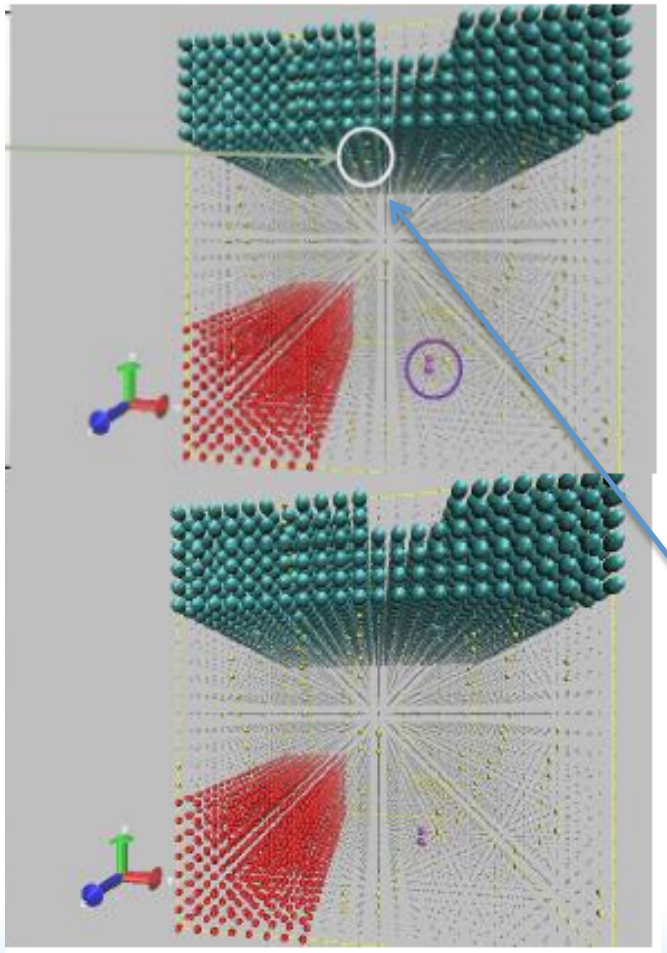
# 5 – More Work Needed - NEIS-P$^2$ – Community Engagement

- Community Outreach and Education activities to enable the general computational science and engineering community to make effective use of petascale systems. This component includes activities to help train the next generation of petascale computational experts through a coordinated **set of courses, workshops, and fellowships.**
- **Key Activities include**
  - **Hands-on workshops**
  - **Virtual School**: semester-long courses on the web to allow participation by students at multiple institutions across the country. The courses will be offered as a traditional college course, including a syllabus with learning outcomes**,**
    - Prototype course taught by Wen-Mei Hwu in Spring 2013.
- **Graduate Fellowships**
  - **Fellowships announced November 11, 2013**
    http://www.ncsa.illinois.edu/news/story/applications_now_being_accepted_for_blue_waters_graduate_fellowships
    - Candidates must already be enrolled in a PhD program at an accredited US non-profit academic institution at the time of application.
    - They must have completed no more than two years of graduate studies. The fellowship support is for one year, renewable based on performance for up to two additional years.
    - The level of support is up to $50,000 per year encompassing a stipend of $38,000 plus $12K in support of tuition and fees as well as support for travel to augment their learning and present papers in their field.
    - Must be US Citizen or Permanent Resident
  - **Internships**
    - Continuing effort from deployment phase
    - Available to undergraduate and graduate students
    - $5K, 1 week hands-on workshop at NCSA
    - Interns are paired with researcher(s)
    - Emphasis on engaging women, minorities and people with disabilities
  - **Blue Waters Symposium**
    - Showcases results from the Blue Waters system, and provides a forum for dealing with community issues and solutions for efficient parallel and heterogeneous petascale computing
    - Planning for this event is underway, but date and location are still under consideration.

# 5 - More Work Needed - NEIS-P2 –Petascale Application Improvement Discovery (PAID) Program

- We are currently revising the details in response to the NSF panel suggestions.
- **Goals**
  - facilitate the creation of new methods and approaches that will dramatically improve the ability to achieve sustained science on petascale systems
  - assist the general computational science community in making effective use of systems at all scales.
- **Major Areas** from Component 1 and Production Experiences
  - Enable application-based topology awareness to more effectively and efficiently use limited bandwidth resources, and to fully exploit the new system functionality for topology aware scheduling that will be available on Blue Waters in 2014.
  - Increasing scalability of full applications, including much work with improving the load balancing within the applications.
  - Improve single node performance for applications, particularly to assist applications in layout, affinity, etc.
  - Increase the number of science applications that can use accelerators and many core technology by lowering the effort to re-engineer applications for these technologies and enabling the teams to maintain a single code base that can be applied to multiple architectures.
  - Enable integrated, at scale applications use of heterogeneous systems that have both general-purpose CPUs and acceleration units.
  - Improve the use of advanced storage and data movement methods to increase the efficiency and time to solution of applications.
  - Assessment and dissemination of science and society impacts resulting from petascale Science

# 6 - Topology Matters – Good and Bad

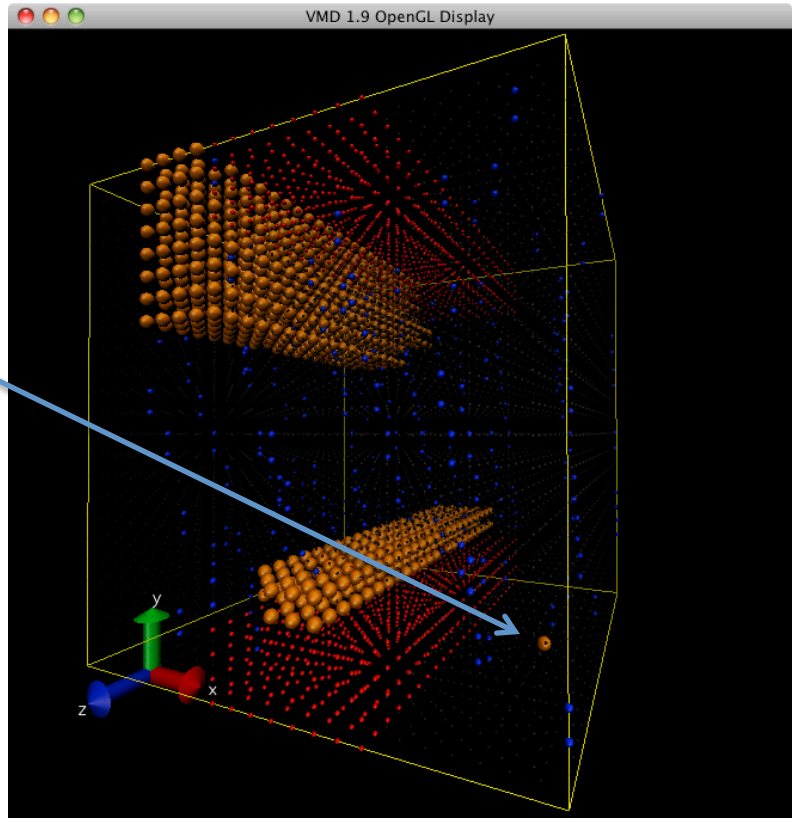


Much of the Benchmark tuning was topology based

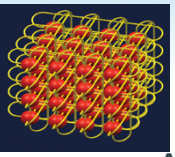
1 poorly placed node out of 4116 (0.02%) can slow an application by >30% (on dedicated system)

Just 1 of 3057 gemini down out in the wrong place of 6114 can slow an application by >20% (P3DNS – 6114 Nodes)

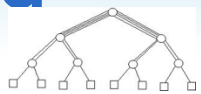
See BW Presentations Later for Positive Impacts

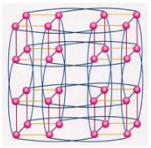



Appears in all system and many applications, but scale makes it clear

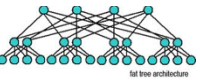


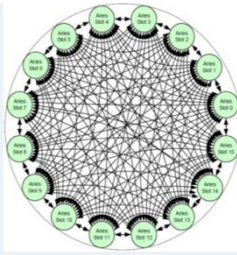
# 6 – Topology Matters - Performance and Scalability through Flexibility

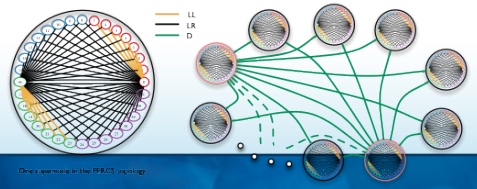




- Applies to all systems and topologies
- Need a system and application partnership to do the best
- Cray developed new management and tuning functions
  - Bandwidth Injection and Congestion Protection features – helps all systems
- BW works with science teams and technology providers to
  - Understand and develop better process-to-node mapping analysis to determine behavior and usage patterns.
  - Better instrumentation of what the network is really doing
  - Topology aware resource and systems management that enable and reward topology aware applications
  - Malleability – for applications and systems
    - Understanding topology given and maximizing effectiveness
    - Being able to express desired topology based on algorithms
    - Mid ware support
- Even if applications scale, consistency becomes an increasing issue for systems applications
- This will only get worse in future systems

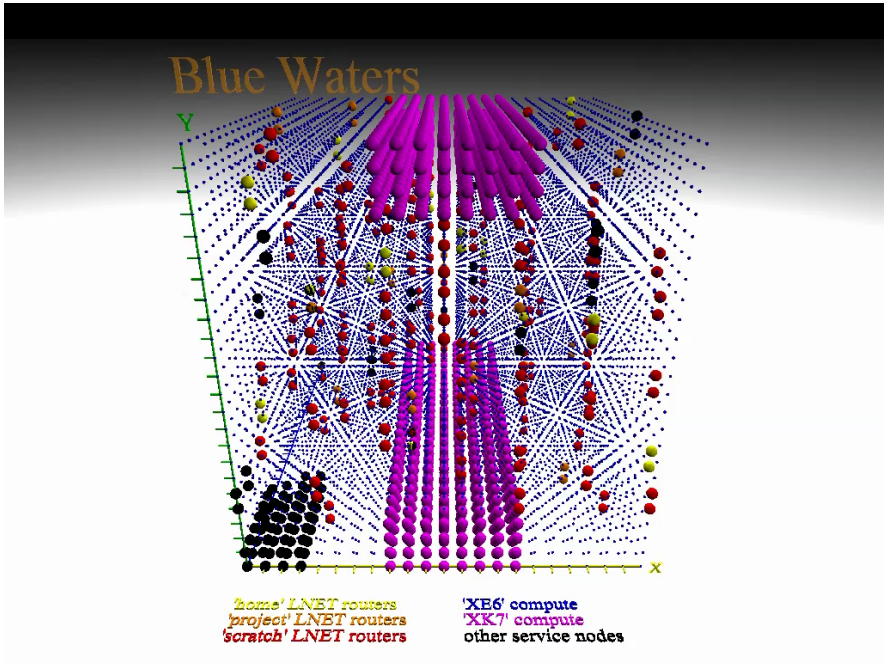






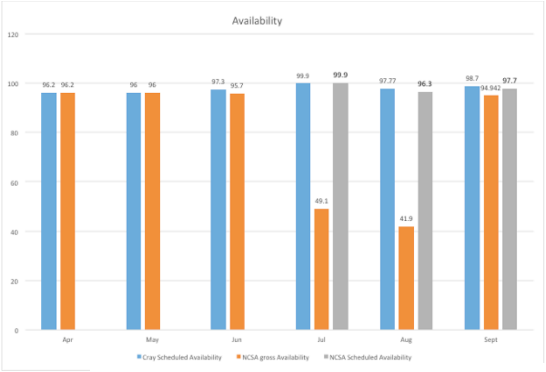
# 7 – I/O thinking is inverted

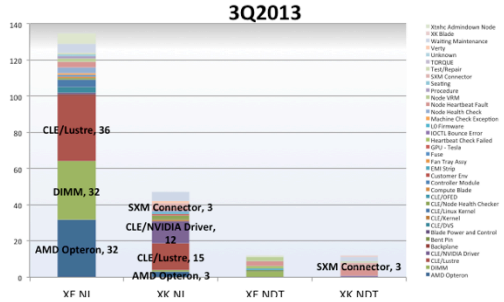
- Many Challenges
  - Scale – timings
  - Failure over works – just not fast enough
- Data and I/O server placement make this a complicated topology based optimization
- Reads slower than writes at scale – one to many rather than many to one
- More in Blue Waters presentation later for more information

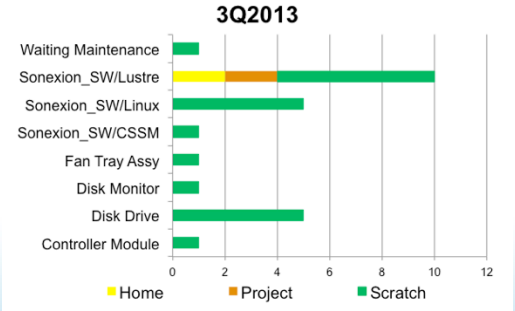


# 8 – Resiliency thinking Inverted

- Current State
  - 2-3 node failures a day – getting less (lower 2's now)
  - SWOs – average 5 days system wide MTBF over several months
- Hardware is better than expected
  - E.g. just 7 HDD failures in 6 months
- Software causes the majority of the failures
  - Almost all storage failures
  - More than ½ the node failures
  - More than ½ the system wide outages
  - Possible the cause of congestion events
- Most research is about the hardware not software
- Most resiliency concerns are about hardware not software



# 8 – Consistency is Getting Worse

- Becoming Intolerable
  - Topology helps – sometimes – but not enough
  - Same code – same nodes – different time
- Most is not OS Jitter related
- Congestion Events Intrusive – see Blue Waters' presentations later for more details
  - Prediction should be possible – but not currently available
- Causes overestimation of run time and less efficient system scheduling
- Impacts resource requirements estimation
- Not just at the largest scales – see slide courtesy of Tom Pugh – Australian Met Office

# 9 - Energy/Heat Control Loops

- 5 independent cooling control loops
- Multi-variants
- 4 orders of magnitude in response times

- Modern Data Center
  - 90,000+ $ft^2$ (8,360+ $m^2$) total
  - 30,000 $ft^2$ (2,790+ $m^2$) raised floor
    20,000 $ft^2$ (1,860+ $m^2$) machine room gallery

- Energy Efficiency
  - LEED certified Gold
  - Power Utilization Efficiency, PUE = 1.1–1.2
  - Staff participating in Energy Efficient HPC working group and

# 10 - Monitoring at Petascale a Big Data Problem

| SOURCE | Average MBs/Day | Max MBs/Day |
|---|---|---|
| apstat | 0.05 | 0.06 |
| bwbackup | 0.06 | 0.74 |
| esms | 229.79 | 622.71 |
| hpss | 345.29 | 1391.80 |
| hpss_core | 0.08 | 0.22 |
| ibswitch | 0.80 | 1.59 |
| jcc | 0.17 | 0.93 |
| moab | 2539.40 | 5678.32 |
| sched | 0.07 | 0.08 |
| SEL | 0.23 | 0.42 |
| sonexion | 326.67 | 870.15 |
| syslog | 12563.31 | 102626.18 |
| torque | 31.66 | 103.78 |
| volkseti | 11.42 | 27.38 |

**Average**

- **15 GBs/day**
- **>88M events/day**
- **> 10,500 defined events**

**Does not include OVIS and Darshan data collection**

- **Will be significant increases at 1 minute resolution for all nodes**

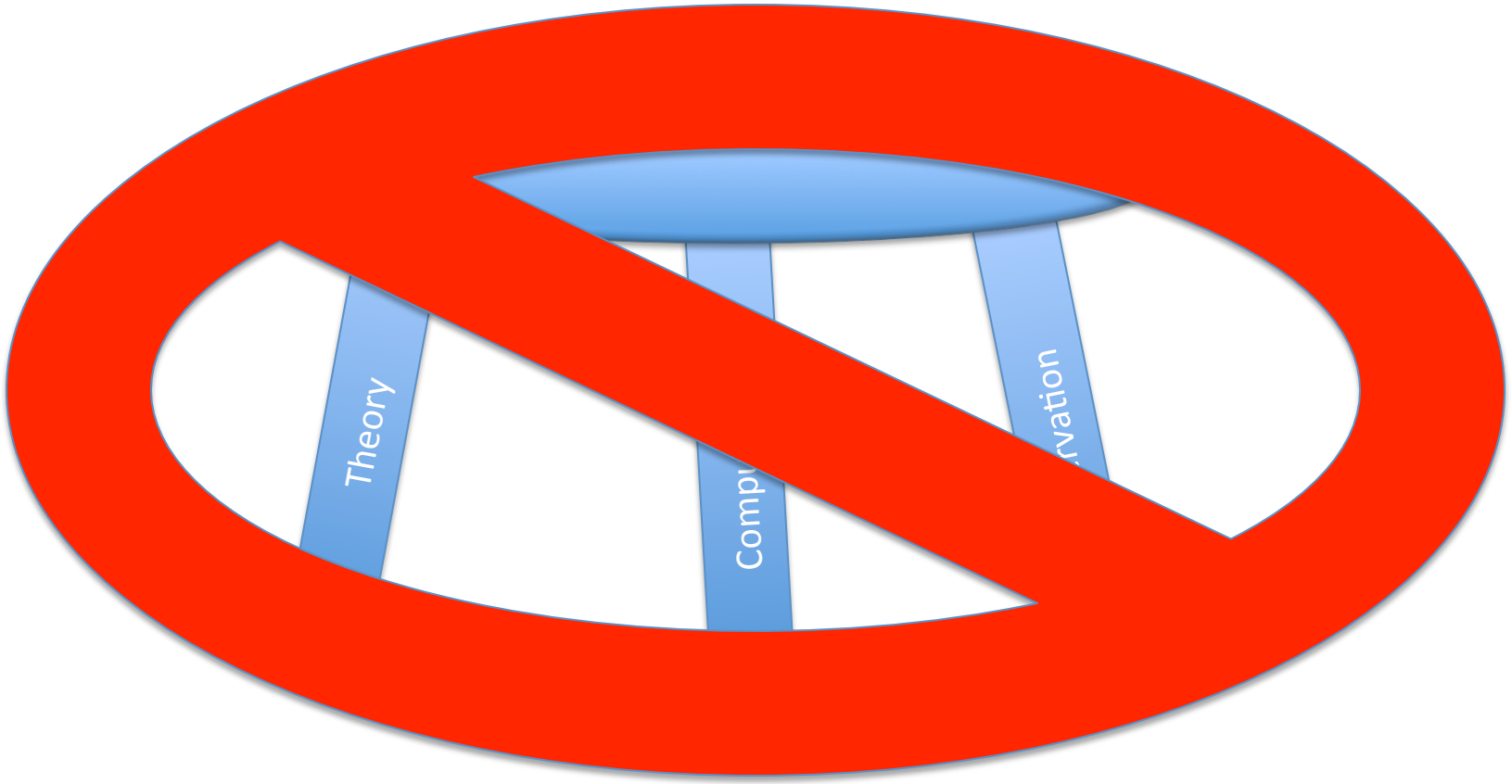
# WHAT COMES NEXT?

Exascale will not be viable until it is Sustained Exascale

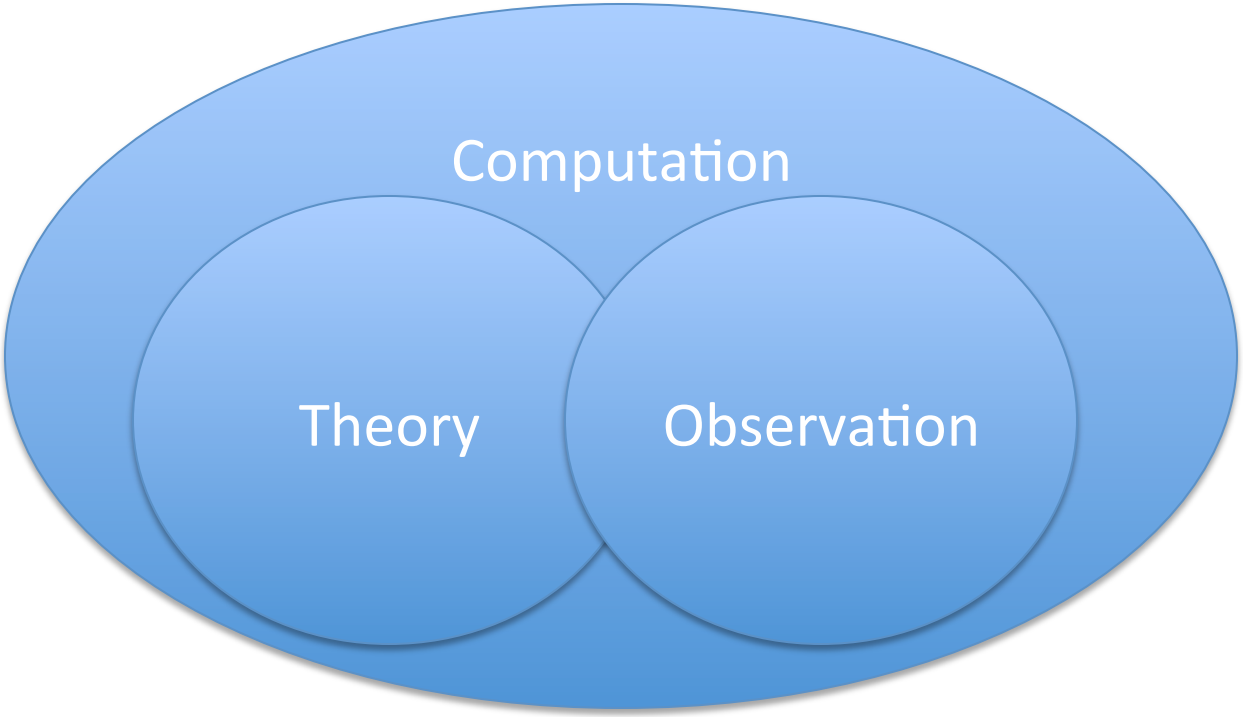
# What Do We Do Now

- Address the issues in Petascale
- No “Killer App” means we have to explain ROI to all science – in only 2 to 3 years
  - Need impact methodology– **quantitative** and qualitative
    - Historical examples is insufficient
- 10/80/10 approach
  - 10% are early adaptors, 10% will never migrated to next new thing
  - Too often we declare success with the 10%
  - The impact comes when the bar is low enough and the real benefit is higher enough for the 80% to move to the new thing
- All we learned about Terascale computing is just a chip - need to reapply and re-implement
- Synergy with Big Data
  - Computer Processed Semi-Structured Data – HPC has been doing this a long time and reasonable well
  - Structured Observational Data – HPC has been doing this a long time and reasonable well
  - Unstructured Observational Data – System capabilities now allow us to consider doing this at unprecedented scale

# Questions



# Questions



# Acknowledgements

This work is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (award number OCI 07-25070) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign, its National Center for Supercomputing Applications, Cray, and the Great Lakes Consortium for Petascale Computation.

The work described is achievable through the efforts of the many other on different teams.