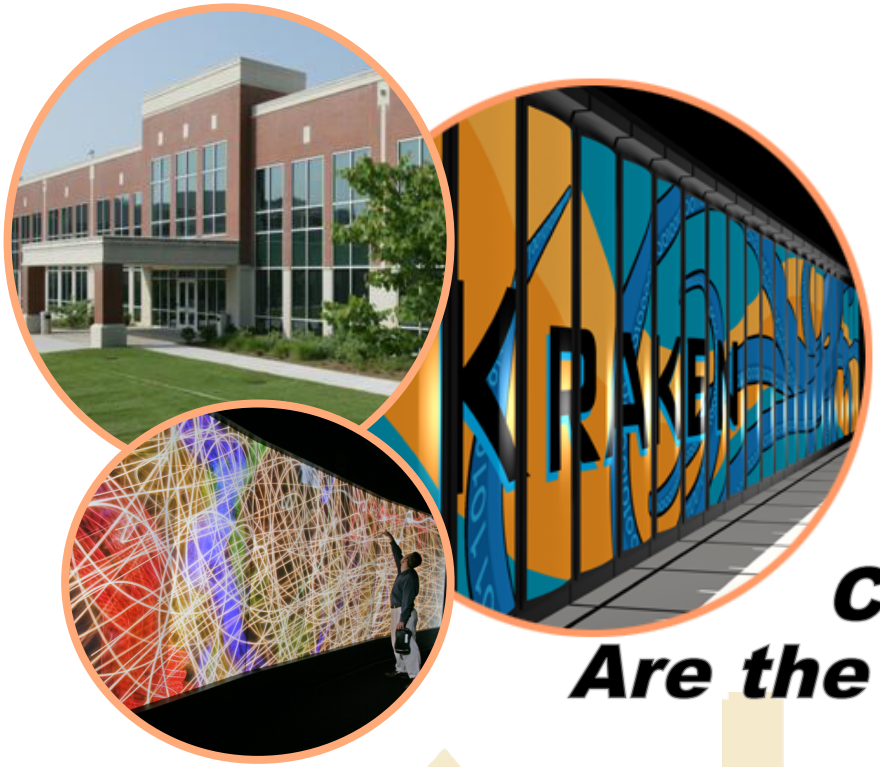


A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

The Malthusian Catastrophe is Upon Us! Are the Largest HPC Machines Ever Up?

Matthew Ezell, Ryan Braby and Patricia Kovatch

University of Tennessee
NICS

The Malthusian Catastrophe

- **Thomas Malthus believed: Human population is growing geometrically, but food production only grows linearly**

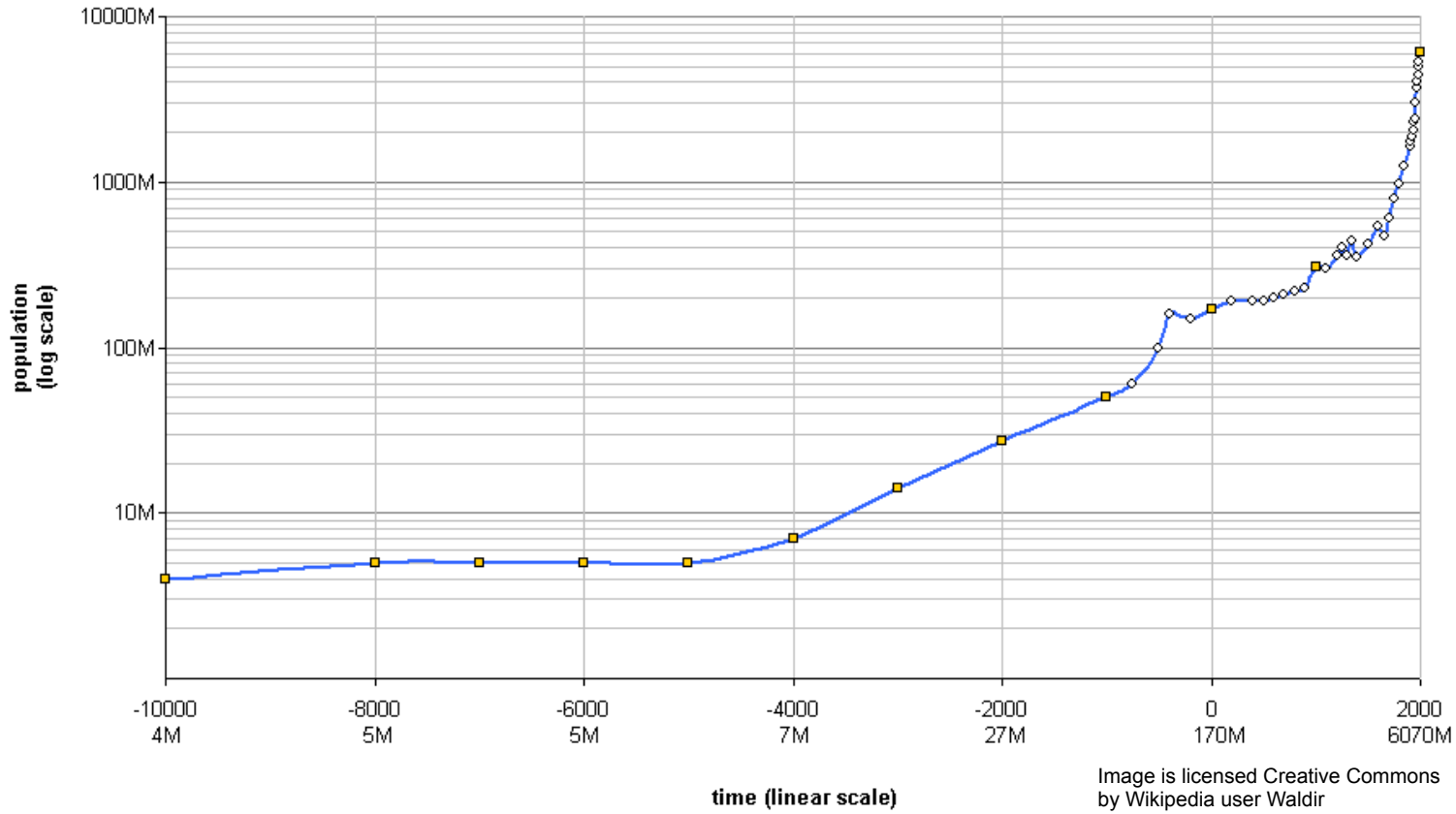


Image is licensed Creative Commons by Wikipedia user Waldir

The Malthusian Catastrophe

- NICS Questions: Supercomputer peak speed is growing geometrically, but are reliability features only improving linearly?

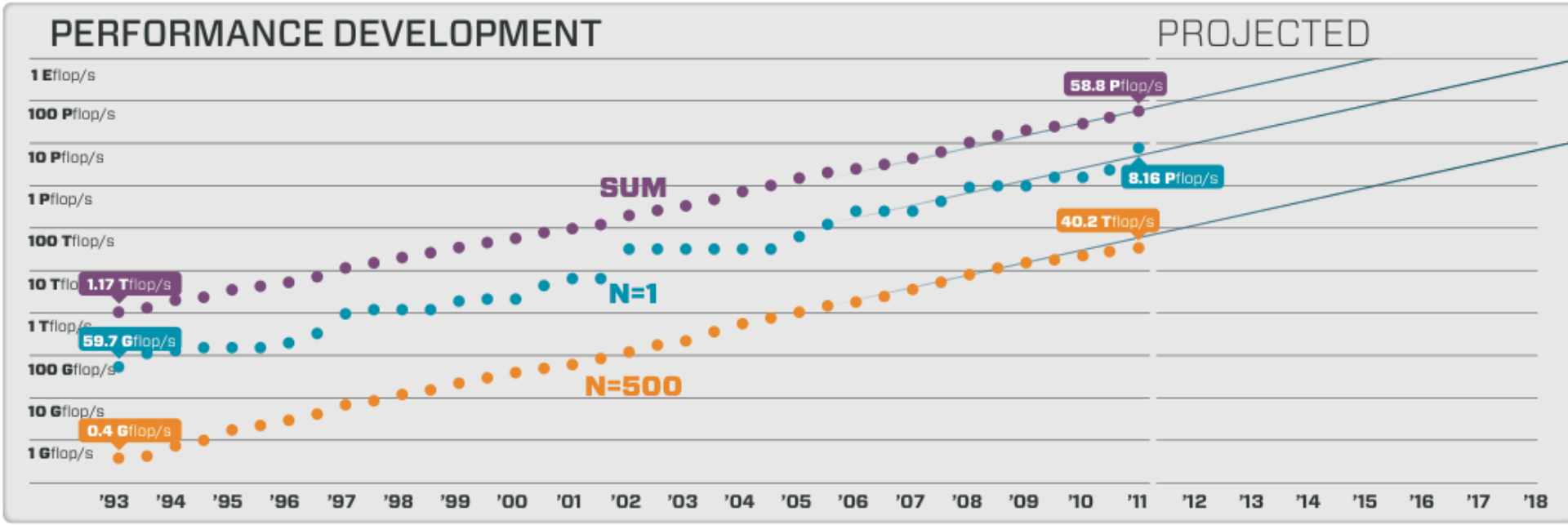
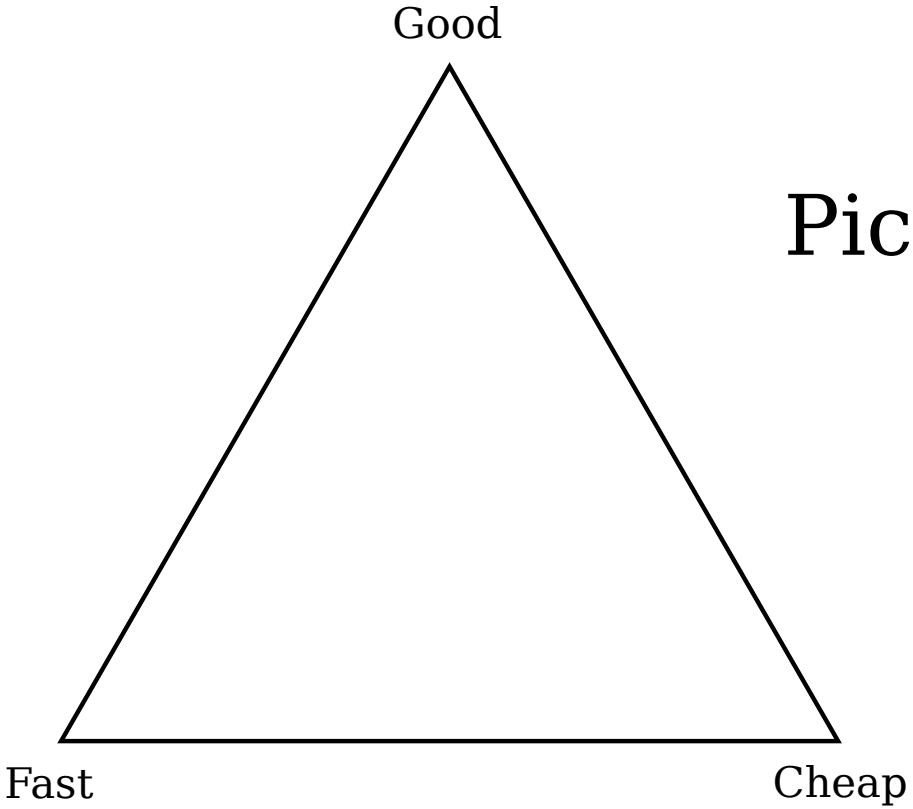


Image obtained from top500.org

Is the sky falling?

The Project Triangle

Performance ⟺ Reliability



Pick any Two

NICS

Defining Resiliency

$$MTBF_{System} = \frac{production\ time}{number\ of\ system\ failures}$$

$$MTBF_{Node} = \frac{production\ time}{number\ of\ node\ failures}$$

Difficulties in Obtaining and Comparing Data

- No common standard for reporting
- It's unclear how to compare different architectures
- Vendors don't like to talk about it
- Root-cause analysis is difficult at best
  - Site staff misdiagnose the failure
  - Vendor retest does not duplicate operating conditions
  - Failure is intermittent and does no fail on retest
  - Proactive replacement

National Institute for Computational Sciences

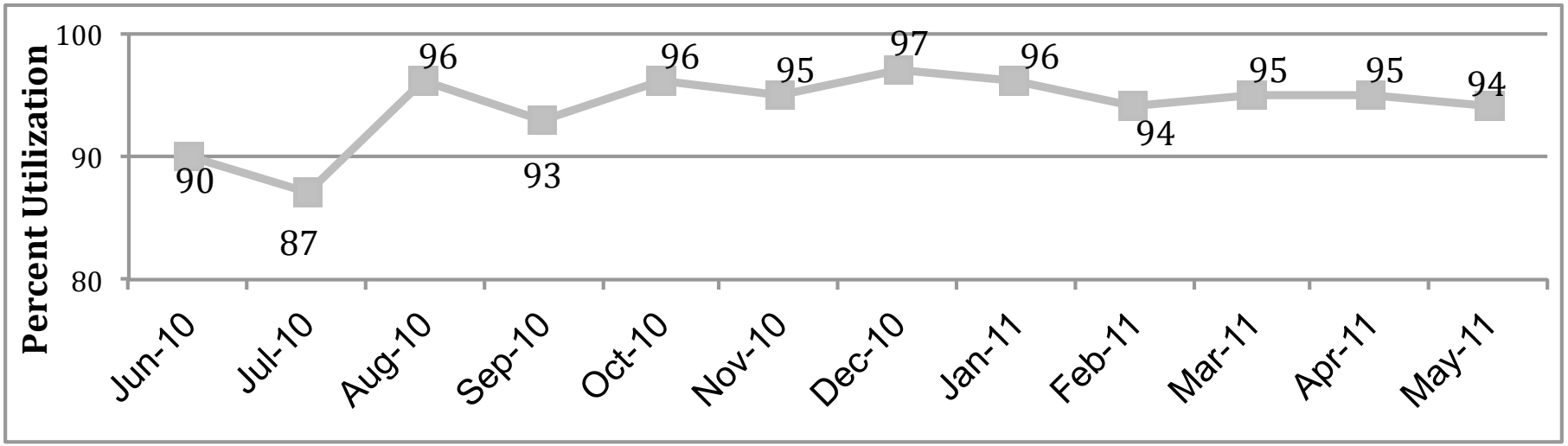
University of Tennessee

- We run three production machines
  - Keeneland (HP GPU Cluster)
  - Nautilus (SGI UltraViolet)
  - Kraken (Cray XT5)

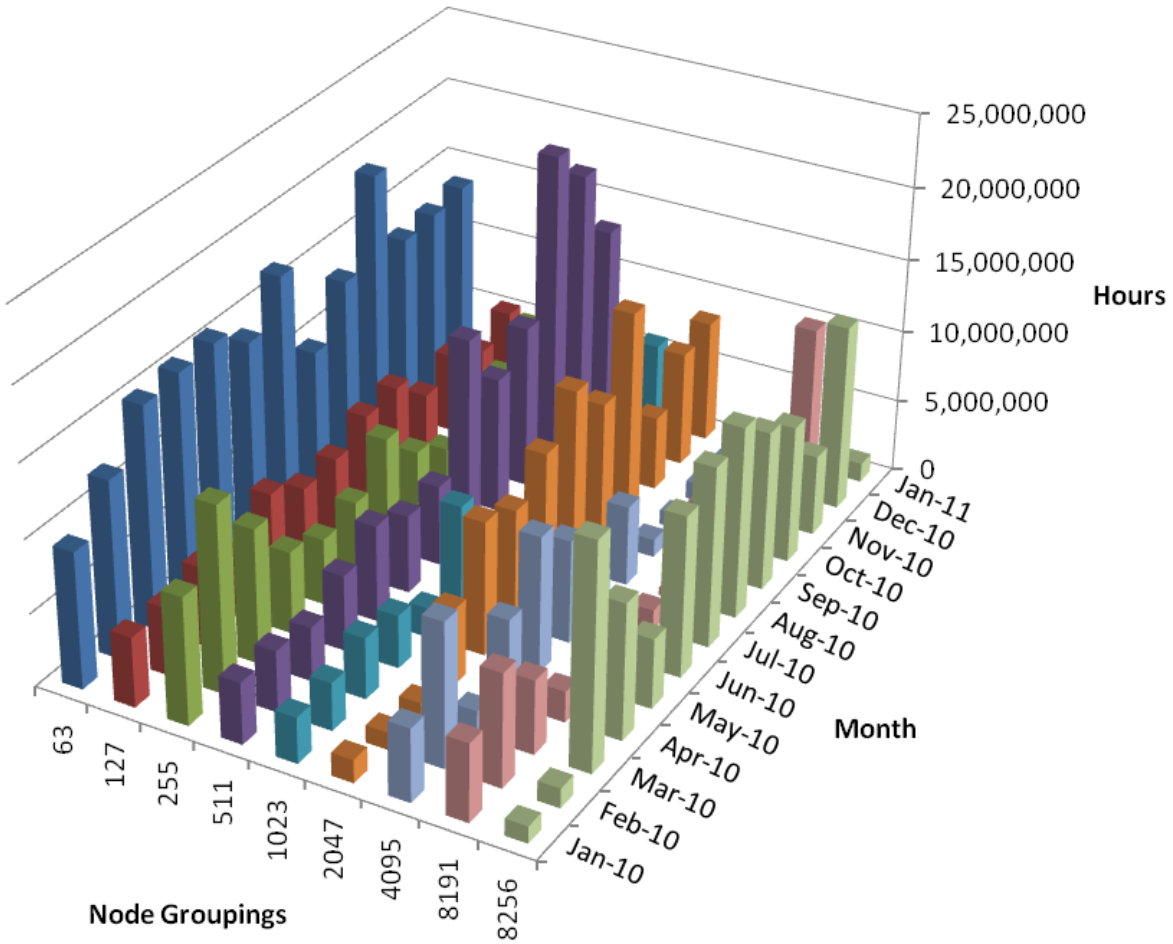
Kraken XT5

Current Kraken Configuration	
Cabinets	100
Interconnect	SeaStar2 3D Torus
Peak Speed	1173 Teraflops
Compute processor type	AMD 2.6 GHz Istanbul-6
Compute cores	112,896
Compute nodes	9,408
Memory per node	16 GB (1.33 GB/core)
Total memory	147 TB

Kraken Utilization over Time



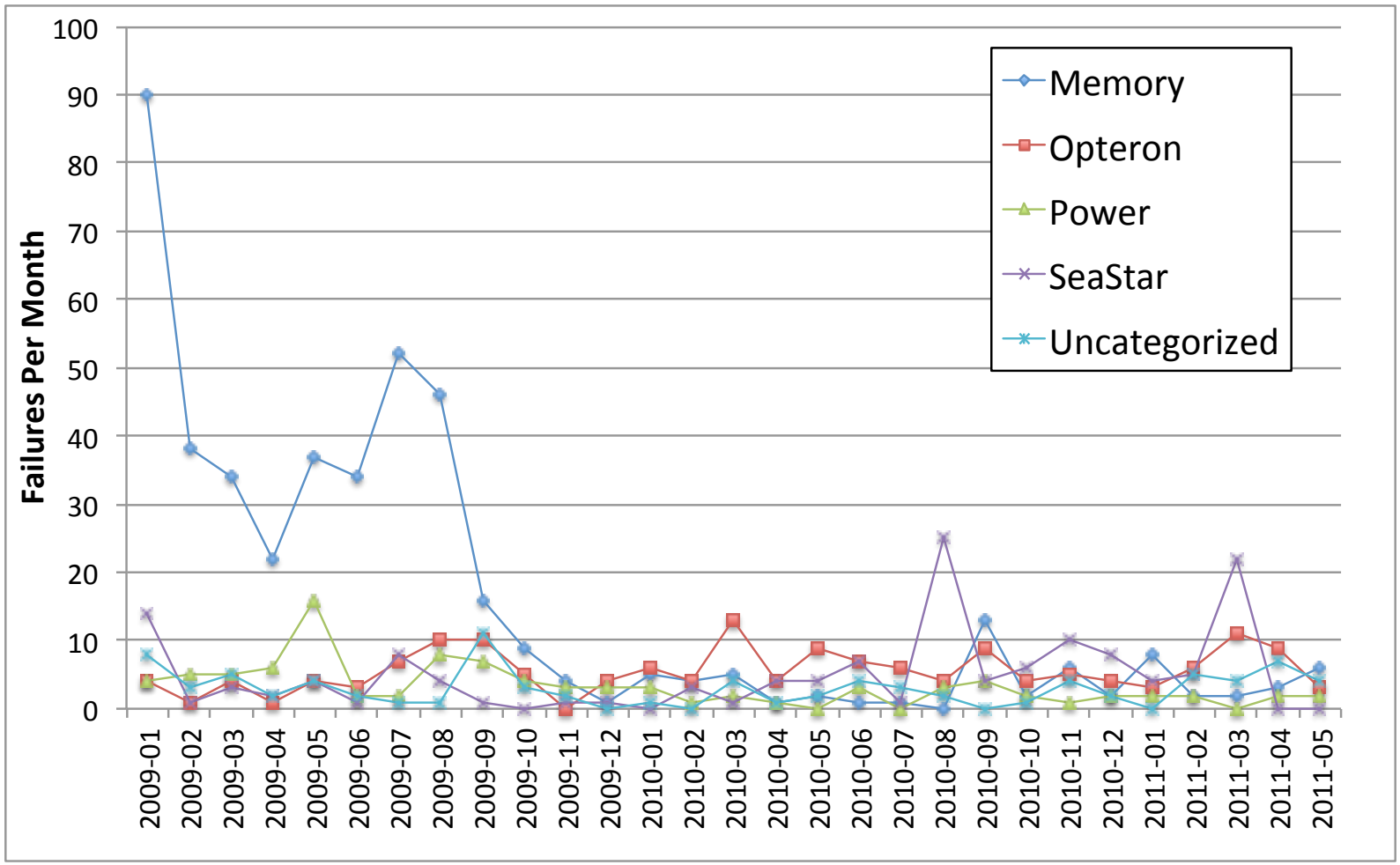
Workload


25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
Hours
Jan-11
Dec-10
Nov-10
Oct-10
Sep-10
Aug-10
Jul-10
Jun-10
May-10
Apr-10
Mar-10
Feb-10
Jan-10
Month
63
127
255
511
1023
2047
4095
8191
8256
Node Groupings

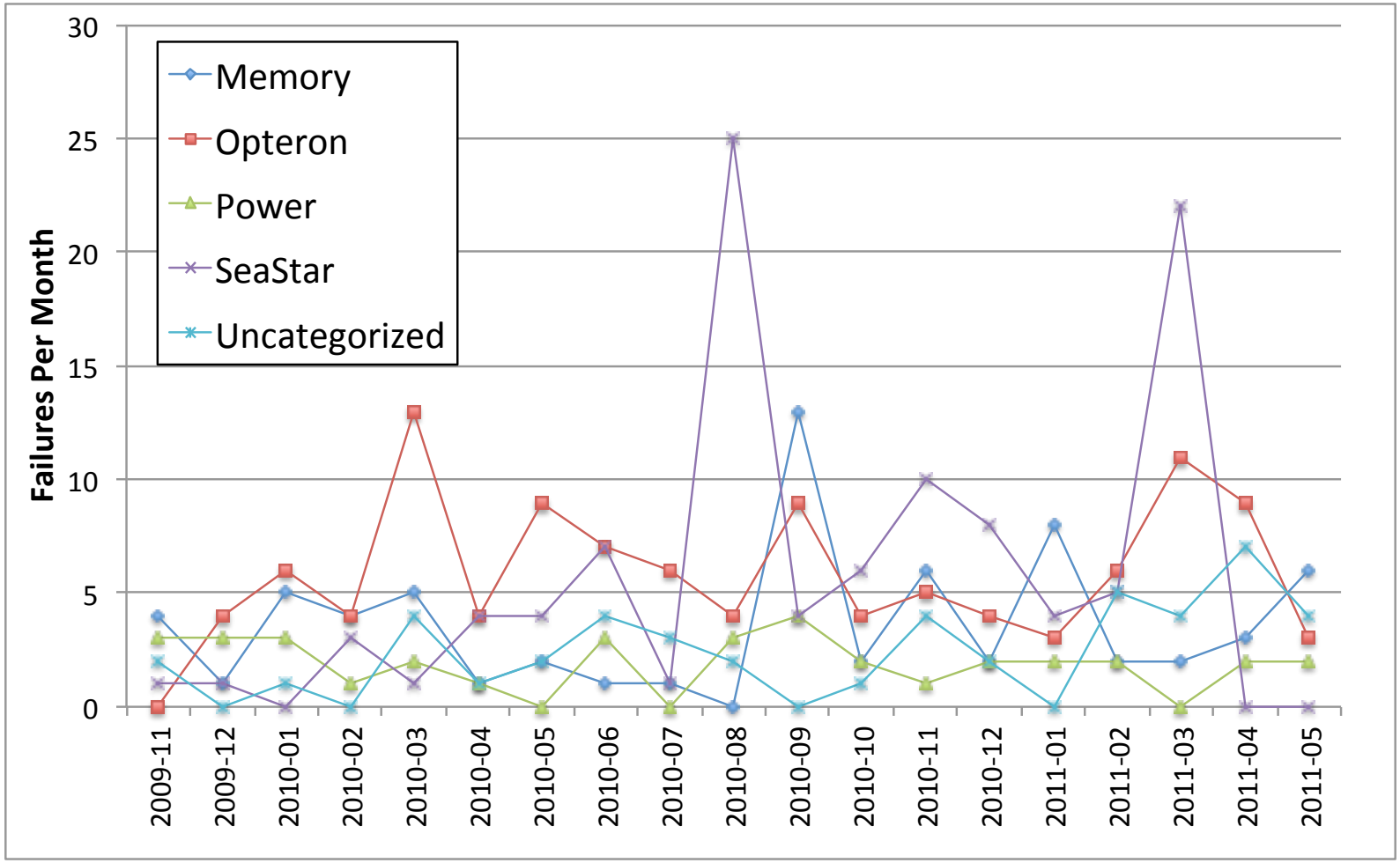
Failure Data Collection at NICS

- Full system outages are manually stored in a database
- Simple Event Correlator (SEC) watches the logs and reports errors to multiple places
  - E-mail to all administrators
  - Log file
  - Cases opened with Cray
- Cray SFDC for other issues
  - Used at every Cray site!

Kraken Node Drop Categories



Kraken Node Drop Categories



NICS

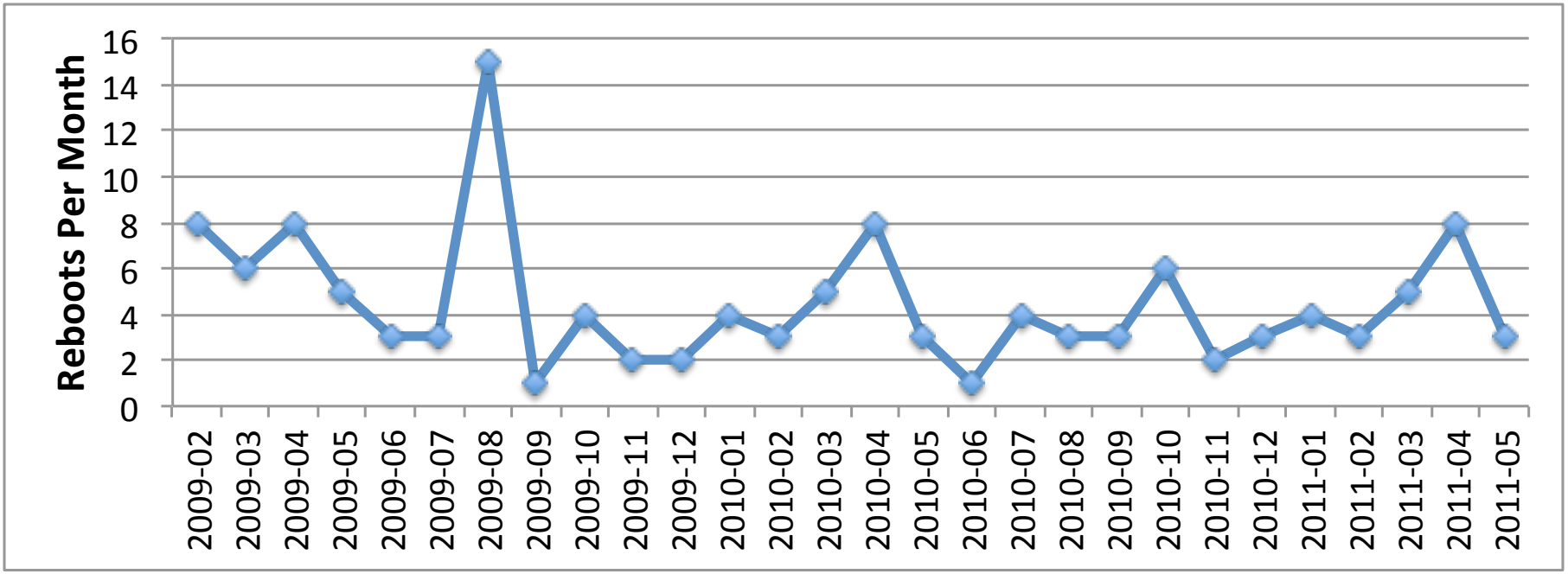
Why the spikes in SeaStar Failures?

- Mostly attributed to CRC errors
- Took out nodes and jobs, but not the entire machine
- A single failing part might cause many errors before it is taken out of service and replaced

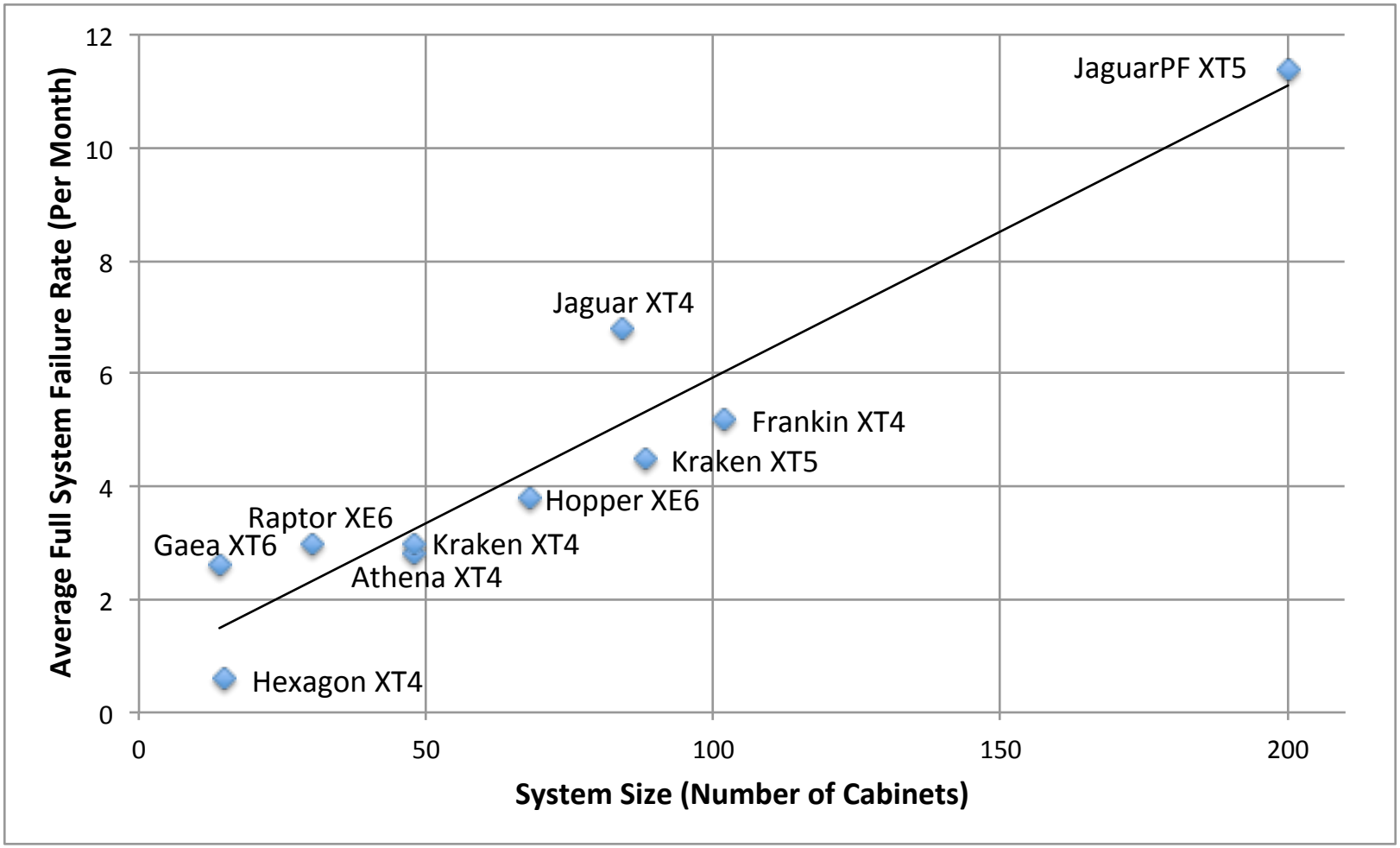
Why so many Opteron Failures?

- Bug in Cray's BIOS
  - Error porting from Barcelona to Istanbul
- Patch received from Cray
  - Might reduce Opteron failures up to 50%

Kraken Full System Failures (Reboots)



Cray Full System Failures



Matt Ezell

University of Tennessee

National Institute for Computational Sciences
ezell@nics.utk.edu