



The Roar of Thunder: LLNL Goes Itanium in a Big Way

Robin Goldstone
Lawrence Livermore National Laboratory

Presented to Gelato.org

May 24, 2004

UCRL-PRES-204277



Background:

HPC Linux Strategy

Platforms Overview

Software Environment



HPC Linux Cluster Strategy

- Create low-cost alternatives to vendor integrated solutions
 - motivates vendors to cut costs
 - provides labs with a level of independence from vendor solutions
- Build clusters from (mostly) commodity components
- Design hardware and software for ease of manageability
- Leverage Open Source software model
- Augment base Linux distribution with in-house development expertise and vendor partnerships
- Develop best of breed software including:
 - Robust, scalable cluster management toolset
 - Efficient, scalable resource manager to exploit maximum utilization of resources
- Employ multi-tiered software support model:
 - LC sysadmins, on-site developers, vendor partners, Open Source community
- Employ “self maintenance” model for hardware repair
- Provide users with a feature-rich environment including:
 - Parallel environment (MPI, OpenMP)
 - Development tools
 - Parallel filesystem
- Deliver world-class HPC systems to our users



LC Linux Cluster Evolution

- PCR, Fall 2001
 - Emperor (88 nodes) and adelic (128 nodes)
 - ia32, dual Xeon, 2GB memory/node, Intel i860 chipset
 - CHAOS 1.1 (RedHat 7.1)
 - Quadrics Elan3 interconnect
- MCR, Fall 2002 (#3 on June 2003 Top500 List)
 - 1152 nodes, dual Xeon, 4GB memory/node, Intel e7500 chipset
 - Federated Elan3 interconnect
 - LinuxBIOS
 - Linux NetworX IceBox for console/power management
 - CHAOS 1.2 (RedHat 7.3)
 - First Lustre deployment
- ALC, Late 2002 (#6 on June 2003 Top500 List)
 - 960 nodes, IBM x-Series dual Xeon, Serverworks GC-LE chipset
 - IBM H8 Service Processor for power management and RAS
- Thunder, Late 2003 (est. #2 on June 2004 Top500 List)
 - 1024 nodes, quad Itanium2, 8 GB memory/node, Intel E8870 chipset
 - Federated Elan4 interconnect
 - CHAOS 2.0 (RHEL 3.0)
 - IPMI for power management



LC Production Resources

System	Manufacturer & Model	Operating System	Interconnect	Nodes	CPUs	Memory (GB)	Peak GFLOP
Unclassified Network (OCF)							47,847
<i>Thunder</i>	<i>Intel Tiger4</i>	<i>CHAOS 2.0</i>	<i>Elan4</i>	<i>1024</i>	<i>4,096</i>	<i>8,192</i>	<i>22,938</i>
<i>MCR</i>	<i>Linux NetworX</i>	<i>CHAOS 1.2</i>	<i>Elan3</i>	<i>1152</i>	<i>2,304</i>	<i>4,608</i>	<i>11,059</i>
<i>ALC</i>	<i>IBM xSeries</i>	<i>CHAOS 1.2</i>	<i>Elan3</i>	<i>960</i>	<i>1,920</i>	<i>3,840</i>	<i>9,216</i>
Frost	IBM SP	AIX 5.1	Colony DS	68	1,088	1,088	1,632
Blue	IBM SP	AIX 5.1	TB3	264	1,056	396	701
TC2K	Compaq SC ES40	Tru64 5.1b	Elan3	128	512	280	683
<i>iLX</i>	<i>Supermicro</i>	<i>CHAOS 1.2</i>	<i>N/A</i>	<i>67</i>	<i>134</i>	<i>268</i>	<i>678</i>
<i>PVC</i>	<i>Supermicro</i>	<i>CHAOS 1.2</i>	<i>Elan3</i>	<i>64</i>	<i>128</i>	<i>128</i>	<i>614</i>
GPS	Compaq GS320/ES4	Tru64 5.1b	N/A	49	160	344	277
Qbert	Digital 8400	Tru64 5.1b	MC 1.5	2	20	24	25
Riptide	SGI Onyx2	Irix 6.5.13f	8 IR2 Pipes	1	48	37	24
Classified Network (SCF)							41,123
White	IBM SP	AIX 5.1	Colony DS	512	8,192	8,192	12,288
<i>Lilac (xEDTV)</i>	<i>IBM xSeries</i>	<i>CHAOS 1.2</i>	<i>Elan3</i>	<i>768</i>	<i>1,536</i>	<i>3,072</i>	<i>9,186</i>
Violet (pEDTV)	IBM p655	AIX	Federation	128	1,024	2,048	6,144
Magenta (pEDTV)	IBM p655	AIX	Federation	128	1,024	2,048	6,144
<i>Adelie</i>	<i>Supermicro</i>	<i>CHAOS 1.2</i>	<i>Elan3</i>	<i>128</i>	<i>256</i>	<i>512</i>	<i>1,434</i>
<i>Emperor</i>	<i>Supermicro</i>	<i>CHAOS 1.2</i>	<i>Elan3</i>	<i>128</i>	<i>256</i>	<i>512</i>	<i>1,434</i>
<i>Ace</i>	<i>Rackable Systems</i>	<i>CHAOS 1.2</i>	<i>N/A</i>	<i>128</i>	<i>256</i>	<i>512</i>	<i>1,434</i>
S (Blue-Pacific)	IBM SP	AIX 5.1	TB3	488	1,952	1,164	1,296
<i>GViz</i>	<i>Rackable Systems</i>	<i>CHAOS 1.2</i>	<i>Elan3</i>	<i>64</i>	<i>128</i>	<i>256</i>	<i>717</i>
Ice	IBM SP	AIX 5.1	Colony DS	28	448	448	672
<i>SC Cluster</i>	<i>Compaq ES40/ES45</i>	<i>Tru64 5.1b</i>	<i>N/A</i>	<i>40</i>	<i>160</i>	<i>384</i>	<i>235</i>
Whitecap	SGI Onyx3800	Irix 6.5.13f	4 IR3 Pipes	1	96	96	77
Tidalwave	SGI Onyx2	Irix 6.5.13f	16 IR2 Pipes	1	64	24	38
Edgewater	SGI Onyx2	Irix 6.5.13f	10 IR2 Pipes	1	40	18	24

Unclassified	
Linux	93%
UNIX	7%
Classified	
Linux	35%
UNIX	65%

Unclassified	
Capability	71%
Capacity	26%
Serial	2%
Visualization	1%
Classified	
Capability	30%
Capacity	64%
Serial	4%
Visualization	2%



CHAOS Model

- CHAOS = Clustered High Availability Operating System
- Augment standard Linux (Red Hat) distribution with kernel modifications and user space tools to support HPC clusters including:
 - Scalable system management and monitoring tools
 - High performance interconnect (Quadrics Elan3/4)
 - Lustre parallel filesystem
 - Resource Management and Control (SLURM)
- Leverage in-house development team to provide rapid turnaround on bug reports and requests for feature enhancements
- Leverage relationships with partners (RedHat, Quadrics, CFS) to achieve production level hardening of key technologies using Open Source software
- Keep kernel mods to a minimum; leverage relationship with RedHat to get our patches accepted into their distribution
- Provide a framework for release management
- Deploy a production quality HPC Linux software stack



What's in CHAOS?

- LLNL-modified kernel
 - CHAOS 1.2 based on RedHat 2.4.18-27 kernel
 - CHAOS 2.0 based on RHEL 3.0 2.4.21 kernel
- genders configuration management tool
- pdsh parallel distributed shell
- conman remote console management
- powerman remote power management
- ganglia (multicast monitoring system) including node up/down detection
- HM (host monitoring) including optimized snmp agent
- munge (scalable intra-cluster authentication)
- mrsh (munge based remote shell)
- YACI (Yet Another Cluster Installer) fast, scalable multicast-based installer
- LinuxBIOS tools and images
- Quadrics software stack for Elan3 (CHAOS 1.2) and Elan4 (CHAOS 2.0)
- SLURM (Simple Linux Utility for Resource Management)
- Crash dump: mcore (CHAOS 1.2) transitioning to Netdump (CHAOS 2.0)
- Thunder specific
 - MCA support in IA64 kernel
 - FreeIPMI support for Intel Tiger nodes



SLURM Overview

- Design goals
 - Simple: Scheduling complexity external to SLURM
 - Open source: GPL
 - Portable: C-language, *autoconf*, general-purpose plugin mechanism
 - Fault-tolerant: For SLURM daemons and its jobs
 - Secure
 - System administrator friendly: Simple configuration file, supports heterogeneous clusters
 - Scalable to thousands of nodes
- **Flexibility via Plug-in mechanism**
 - Authentication
 - Authd, Munge or none
 - Interconnect
 - Quadrics Elan3, Elan4, Ethernet or Myrinet
 - Under Development: InfiniBand and IBM Federation
 - Scheduler
 - Maui, FIFO or backfill, LSF under consideration
- **Managing ~400 Linux clusters worldwide**

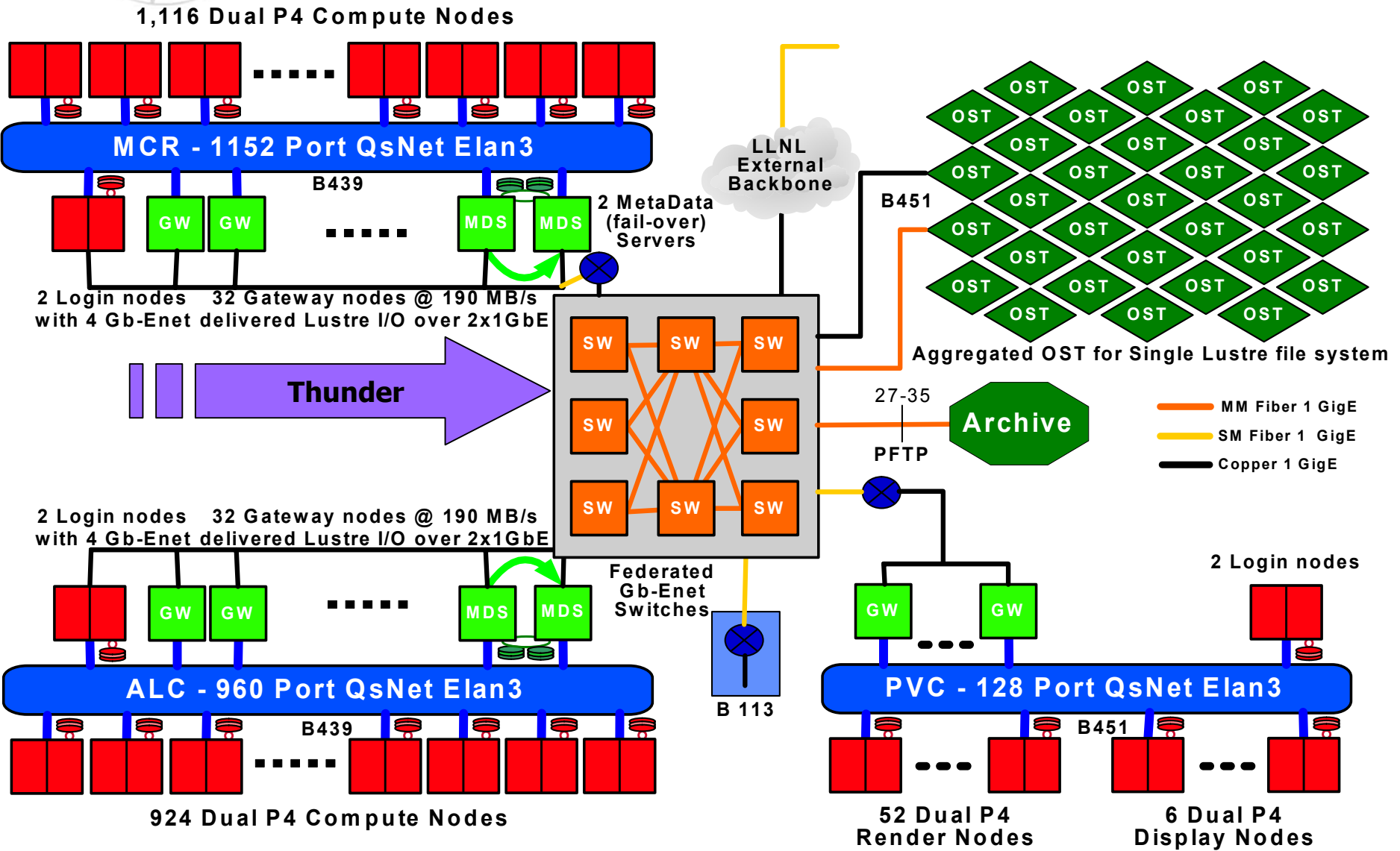


Commitment to Lustre

Cluster	Arch	Nodes	Cpus	Mem/node	OST/GW nodes	Lustre disk space
MCR	ia32	1152	2304	4GB	16OST/ 16GW	173TB
PVC	ia32	70	140	2GB	4GW	Shares MCR file systems
ALC	ia32	960	1920	4GB	32GW	75TB
THUNDER	ia64	1024	4096	8GB	16GW	200TB
LILAC	ia32	768	1536	4GB	32	140TB



The InterGalactic Filesystem





Development Environment

- Compilers
 - Intel compilers (ifort, icc) - through v8
 - PGI compilers (pgcc, pgCC, pgf90, pgHPF) - through 5.1
 - gcc (2.96, 3.2.3, 3.3.2), g77, glibc 2.2.5, 2.3.2
- Libraries
 - Intel MKL (math kernel library) - current is v6.1
 - Other libs (ScaLAPACK, pact, etc.) are responsibility of apps teams
- Interpreters
 - Java - j2re1.4.1_03
 - Python 2.2
- Debuggers
 - TotalView 6 - port with ASC enhancements
 - gdb, ddd, idb, pgdbg
 - ddt - under evaluation
- Profiling
 - mpiP
 - gprof, pgprof
 - Vampir/Guideview
 - PAPI
 - In-house HWmon tool - layered on perfctr
 - TAU (experimental)
 - VTune coming (not yet available for RHEL 3)
- Memory debuggers
 - Valgrind
 - dmalloc, efence, Purify for gcc



Thunder

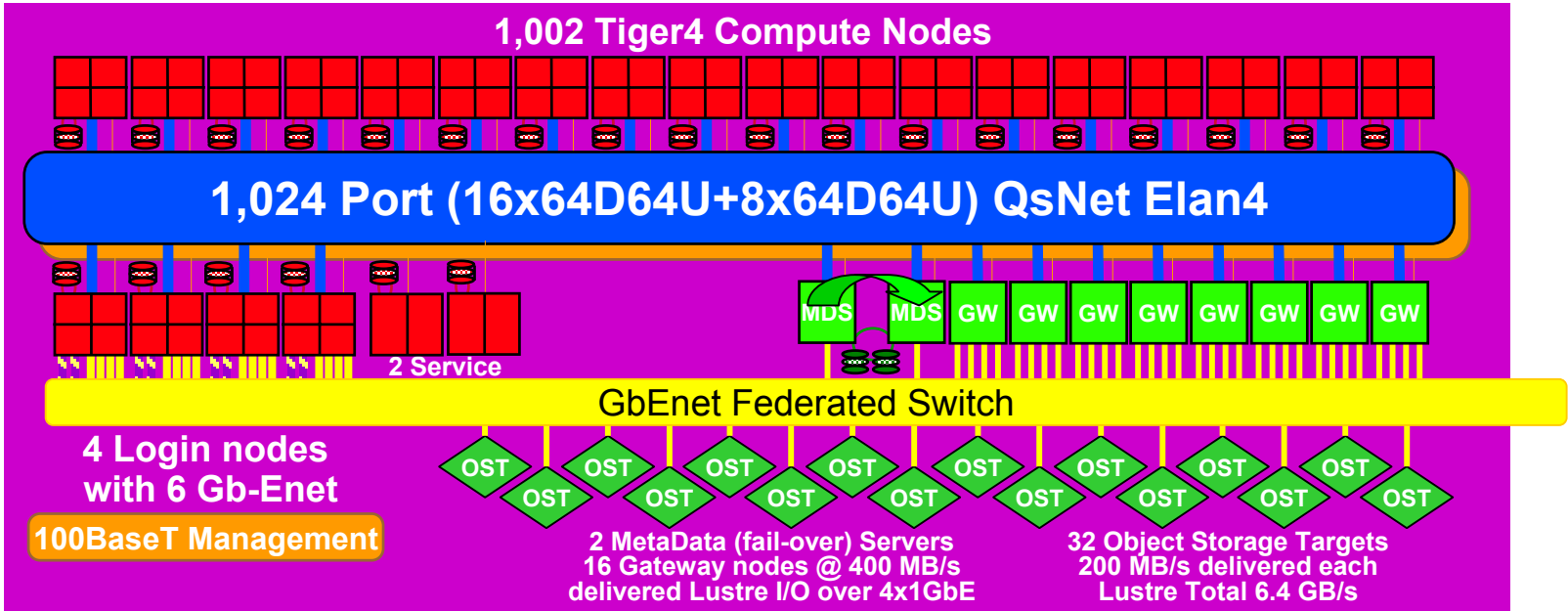


Thunder is...

- A 1024-node, 4096 processor Itanium2 Linux cluster with 8TB memory, 200TB global (Lustre) disk and a peak performance of 22.9TF.
- A collaboration between:
 - LLNL: CHAOS, cluster tools, SLURM
 - Intel: tiger4 nodes, compilers
 - RedHat: base Linux distribution
 - CFS: Lustre filesystem
 - Quadrics: Elan4 interconnect, MPI software stack
 - California Digital: integration, FreeIPMI software
- The fastest supercomputer in North America, as well as having the highest efficiency (87%) of any non-vector architecture (based on HPC Linpack result).



Thunder Architecture



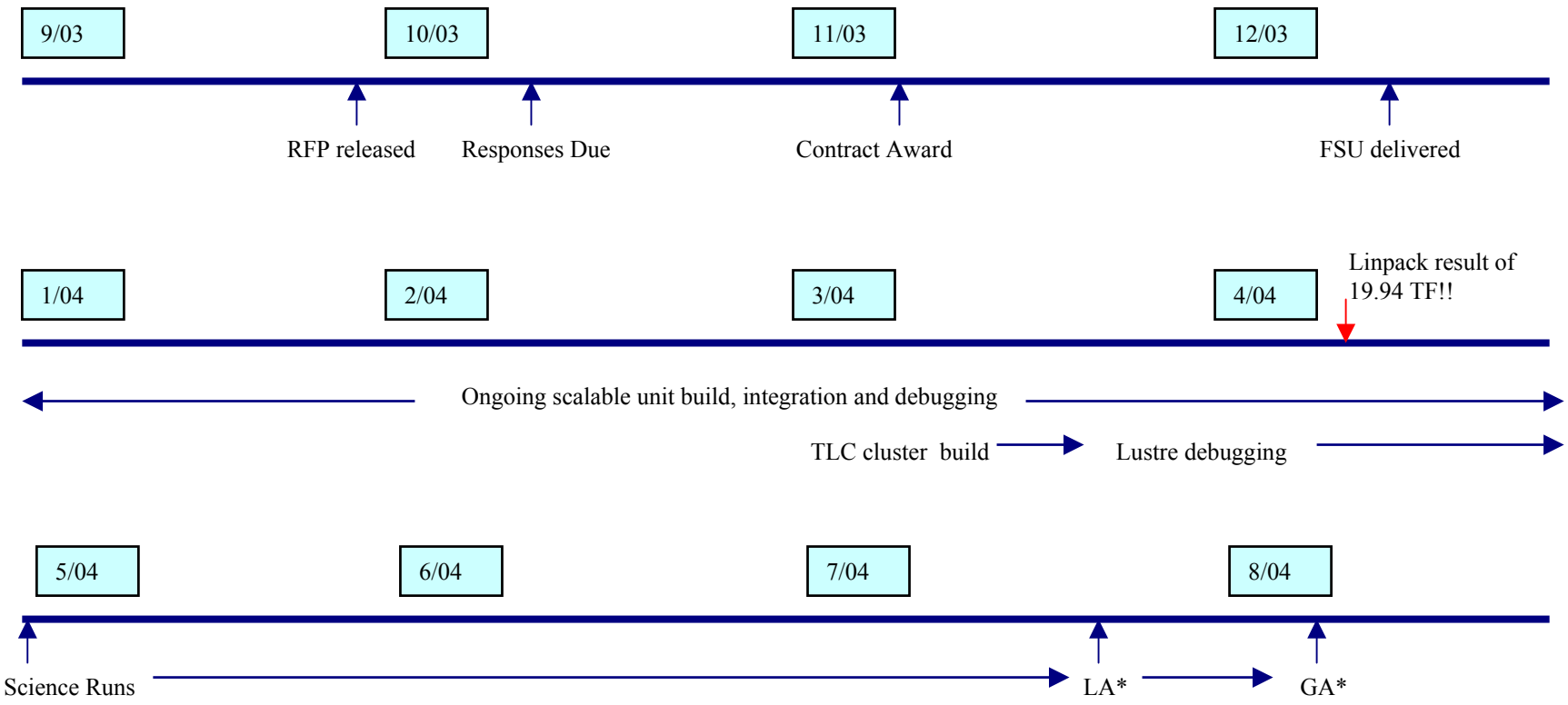
System Parameters

- Quad 1.4 GHz Itanium2 Madison Tiger4 nodes with 8.0 GB DDR266 SDRAM
- QsNet Elan4 with 900 MB/s MPI bandwidth and <3 μs latency
- Support 400 MB/s transfers to Archive over quad Jumbo Frame GigE from each login node
- 75 TB local disk in 73 GB/node UltraSCSI320 disk
- 50 MB/s POSIX serial I/O to any file system
- 20 B:F = 200 TB global storage (DDN fiberchannel/SATA), 6.4 GB/s delivered parallel I/O performance
- Software: CHAOS 2.0, SLURM/LCRM, MPICH2, TotalView, Intel/GNU/PGI compilers

19.94 TF/s on Linpack (~87% of peak)



Thunder Integration Timeline



* Tentative



Thunder-specific software development

- MCA enhancements
 - Rewrite of salinfod for improved robustness and functionality
 - sysctl options to panic at CPE/CMC thresholds and turn console level logging on/off
 - Software scrubbing of correctable memory errors
- FreeIPMI (co-developed with CDC)
 - libfreeipmi framework + FISH
 - ipmipower
 - bmc configuration/query tool
 - bmc watchdog
 - sensor query tool
 - system event log (SEL) view/update
- Misc tiger4 tools
 - CMOS parameter manipulation
 - EFI-based firmware flash



For more info...

- Thunder: www.llnl.gov/linux/thunder
- CHAOS Overview: www.llnl.gov/linux/chaos
- LLNL cluster tools: www.llnl.gov/linux/downloads.html
- SLURM: www.llnl.gov/linux/slurm/slurm.html
- FreeIPMI: <http://www.nongnu.org/freeipmi/>