UCRL-TR-224276

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Report to the Institutional Computing Executive Group (ICEG) August 14, 2006

B. Carnes

September 8, 2006

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Report to the Institutional Computing Executive Group (ICEG)

August 14, 2006

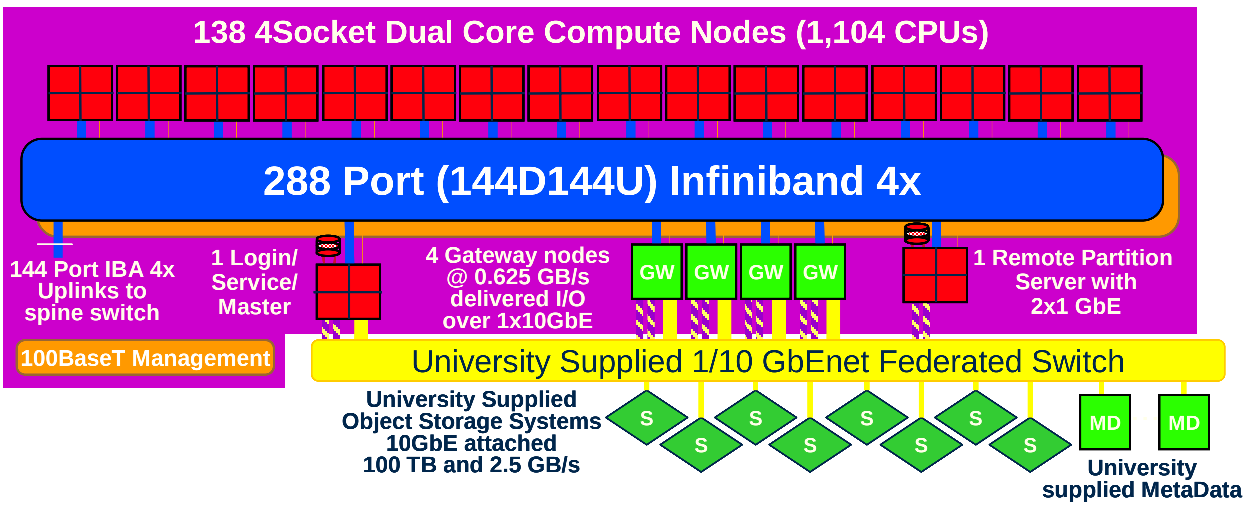


Figure 1.0.1. Peloton Scalable Unit—New M&IC Platform Building Block

1.0 Executive Summary

We have delayed this report from its normal distribution schedule for two reasons. First, due to the coverage provided in the White Paper on Institutional Capability Computing Requirements distributed in August 2005, we felt a separate 2005 ICEG report would not be value added. Second, we wished to provide some specific information about the Peloton procurement and we have just now reached a point in the process where we can make some definitive statements. The Peloton procurement will result in an almost complete replacement of current M&IC systems. We have plans to retire MCR, iLX, and GPS. We will replace them with new parallel and serial capacity systems based on the same node architecture (**Fig. 1.0.1**) in the new Peloton capability system named ATLAS. We are currently adding the first users to the Green Data Oasis, a large file system on the open network that will provide the institution with external collaboration data sharing. Only Thunder will remain from the current M&IC system list and it will be converted from Capability to Capacity. We are confident that we are entering a challenging yet rewarding new phase for the M&IC program.

2.0 Institutional and Programmatic Funding

M&IC is supported by three sources of income (**Fig. 2.0.1**):

- On-going or base G&A funding. This funding provides the necessary support to operate M&IC on a day-to-day basis: staff, power, facility costs, maintenance, SW contracts, plus a basal funding level for new investments in computing capacity.
- Investments from participating programs and directorates. This funding is used to provide access to the programs that co-invest in the hardware. These investments are applied to new procurement lease-to-ownership (LTO[1]s) and other existing LTOs. They also support procurement of visualization, archive, and other infrastructure to maintain a robust and balanced environment.
- A supplemental G&A funding source (incremental—one timers). This institutional support is to cover the LTO costs of new parallel Capability and Capacity systems.

M&IC FY00-FY09 Institutional Funding

Costs	FY00	FY01	FY02	FY03	FY04	FY05	FY06	FY07	FY08	FY09
Base*	2,044	2,055	3,055	4,383	4,208	5,360	5,003	5,203	5,411	5,628
Incremental - One Timers	1,684	3,417	3,300	7,118	9,062	1,158	2,500	7,400	7,800	1,800
Program Contributions	935	935	935	935	1,052	1,023	976	1,023	1,023	1,023
Total Costs	**4,663**	**6,407**	**7,290**	**12,436**	**14,322**	**7,541**	**8,479**	**13,626**	**14,234**	**8,451**

*Out-year increase is an estimate to cover fixed costs

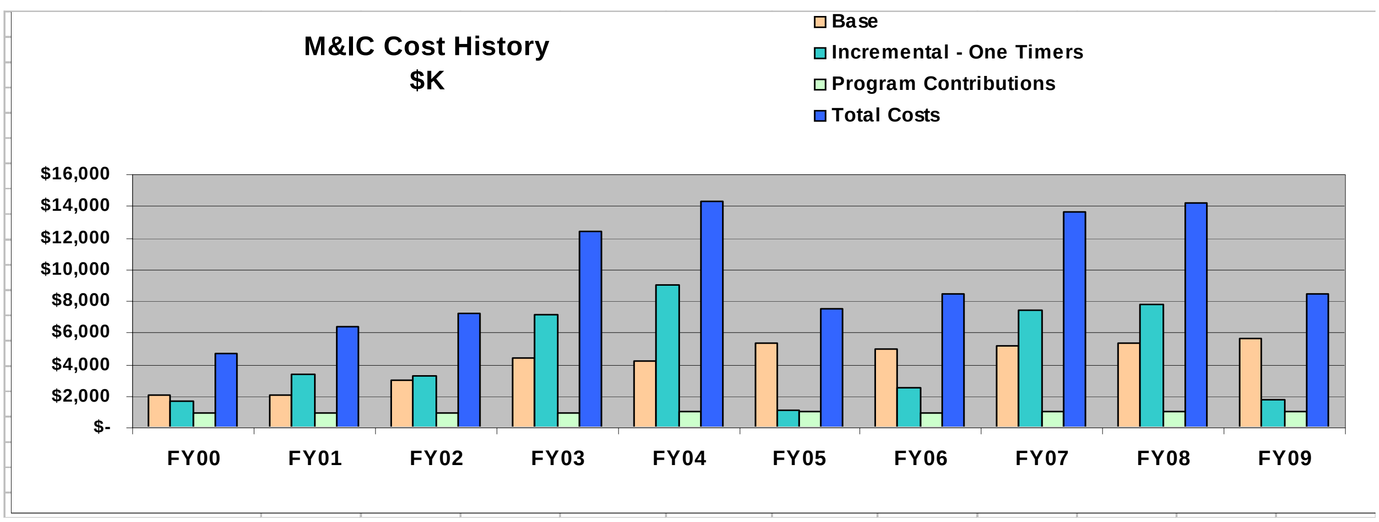


Figure 2.0.1 M&IC Funding History ($K)

[1] Program and directorate funding is volatile and cannot be relied upon with confidence for planning procurements.

2.1 Institutional and Programmatic Allocations

Figure 2.1.1 displays the peak speeds of systems made available in whole or in part to M&IC since FY00 in log scale. In FY07, there will be a total of 81 TERAFLOPS available to the M&IC program, including both Capability and Capacity systems.

The current allocations for investing programs are shown in **Table 2.1.1**. For each M&IC system, we are providing a 30% buy-in bonus (for each dollar pledged, the program receives $1.30 in ownership rights in the system). The institution is intentionally providing high-performance computing for the programs at a very low cost. **Table 2.1.2** shows the history of M&IC systems from FY97 to CY07.

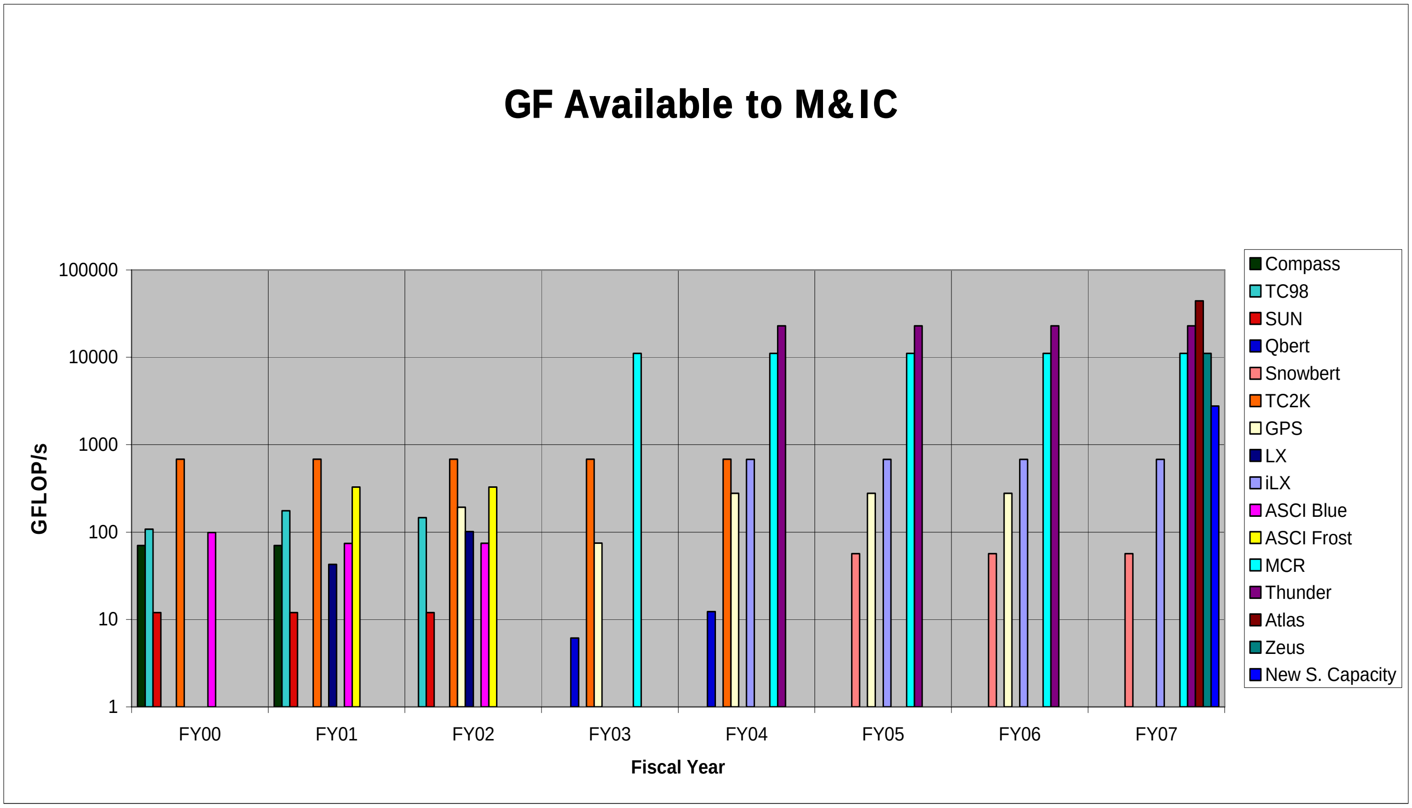


Figure 2.1.1 GFLOPS provided by each system to M&IC users (log scale)

Program Allocations													
Investor	Bank	GPS/ILX				MCR				Thunder			
		Buy-In ($K)	Allocation	CPU-hrs per week	CPUs	Buy-In ($K)	Allocation	CPU-hrs per week	CPUs	Buy-In ($K)	Allocation	CPU-hrs per week	CPUs
Institution	ic	1859	42.28%	21909	130	9614	51.30%	191669	1141	10315	63.60%	428277	2549
D&NT	ds	400	16.19%	7710	46	1415	17.95%	67066	399	1200	10.89%	73328	436
Physics	micphys	317	10.64%	5065	30	625	6.74%	25200	150	625	5.67%	38191	227
CMS	cms	267	10.06%	4790	29	775	8.44%	31517	188	450	4.08%	27498	164
Biosciences	biomed	12	0.30%	142	1					99	0.90%	6050	36
E&E	ees	80	2.92%	1388	8	210	2.27%	8467	50	250	2.27%	15277	91
Lasers/NIF	nif		1.98%	944	6								
Engineering	engr	180	5.96%	2840	17	338	3.64%	13608	81	338	3.06%	20623	123
Comp	casc	33	3.96%	1885	11	710	7.66%	28627	170	100	0.91%	6111	36
DHS	dhs		0.95%	451	3	214	2.00%	7478	45				
UCRP	ucrp		1.04%	494	3								
Q division - MNT	mnt									99	0.90%	6050	36
Pu aging	puage									400	3.63%	24443	145
CIAC	ciac	150	3.73%	1774	11								

Table 2.1.1 Investor Ownership in M&IC resources

Theoretical Peak Performance (GFLOPS)

	FY97	FY98	FY99	FY00	FY01	FY02	FY03	FY04	FY05	FY06	CY07
T3D	37	0	0	0	0	0	0	0	0	0	0
Compass	35	70	70	70	70	0	0	0	0	0	0
TC98	0	0	96	108	176	147	0	0	0	0	0
SUN	0	12	24	12	12	12	0	0	0	0	0
Qbert	0	0	0	0	0	0	12	12	0	0	0
Snowbert	0	0	0	0	0	0	0	0	57	57	57
TC2K	0	0	0	683	683	683	683	683	0	0	0
GPS	0	0	0	0	0	192	277	277	277	277	0
LX	0	0	0	0	43	101	0	0	0	0	0
iLX	0	0	0	0	0	0	634	678	678	678	0
ASCI Blue	0	16	89	99	74	74	0	0	0	0	0
ASCI Frost	0	0	0	0	326	326	0	0	0	0	0
MCR	0	0	0	0	0	11059	11059	11059	11059	11059	0
Thunder	0	0	0	0	0	0	0	22938	22938	22938	22938
Atlas	0	0	0	0	0	0	0	0	0	0	44237
Zeus	0	0	0	0	0	0	0	0	0	0	11059
New S. Capacity	0	0	0	0	0	0	0	0	0	0	2760
Total Peak GF	72	98	279	972	1384	12594	12665	35647	35009	35009	81051

Table 2.1.2 History of M&IC Systems

2.2 Capability System Resource Utilization

Figure 2.2.1 shows Thunder utilization by directorate as a percentage of the available cycles. We expect 15% idle time for a heavily contended system, given the inefficiencies of scheduling mechanisms on the Capability systems; Thunder does somewhat better than that. Institutional usage is broken down by project for Thunder in **Figure 2.2.2**.

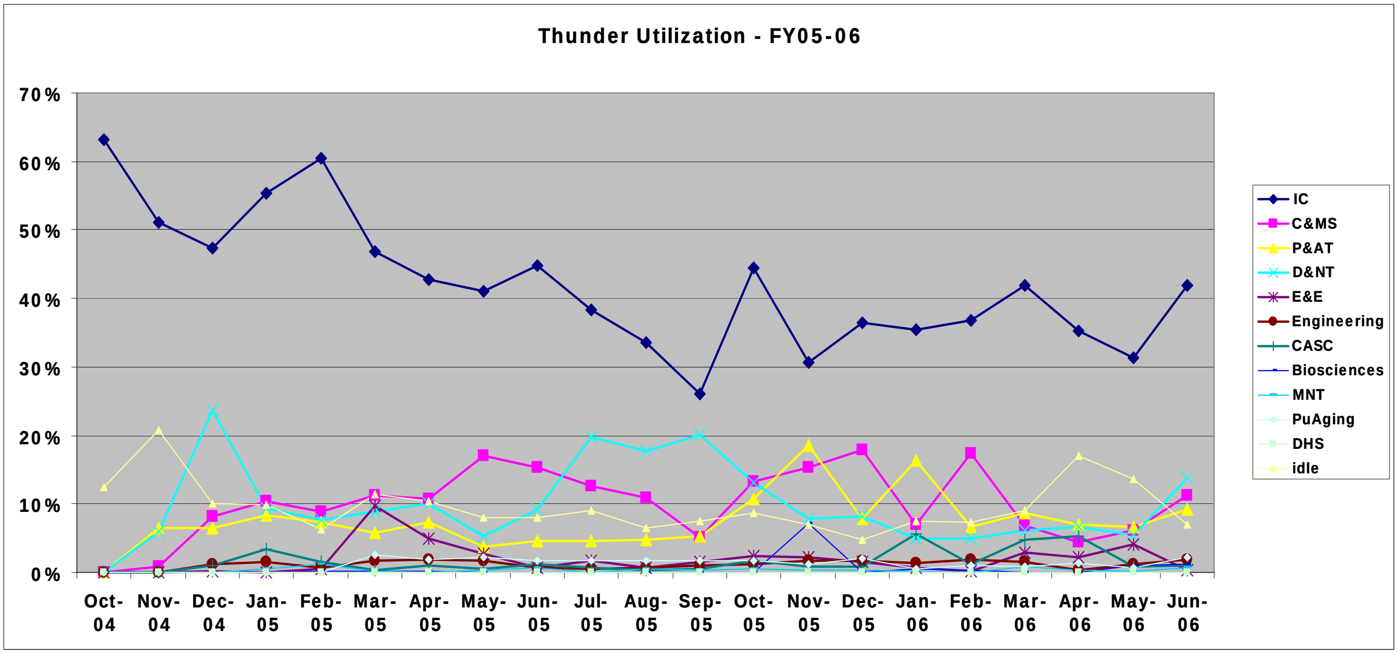


Figure 2.2.1 Institutional utilization of Thunder

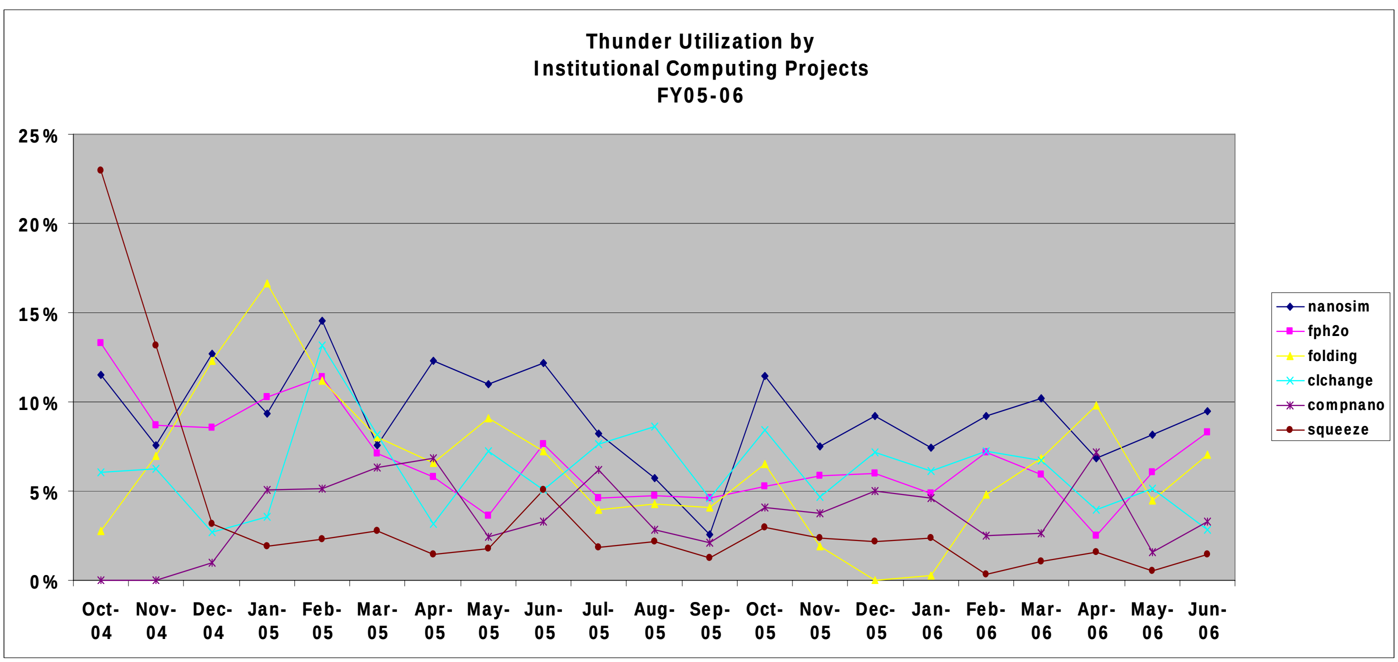


Figure 2.2.2 Institutional utilization of Thunder by project name

2.3 Capacity System Resource Utilization

Capacity systems are intended to serve as a large pool of available cycles on demand for code development and small parallel calculations. We attempt to have enough capacity cycles so that access is available on demand during the day. A scientist should not have to wait for a processor or two to do development and debugging. **Figures 2.3.1** and **2.3.2** show MCR utilization. **Figures 2.3.3** and **2.3.4** show GPS utilization.

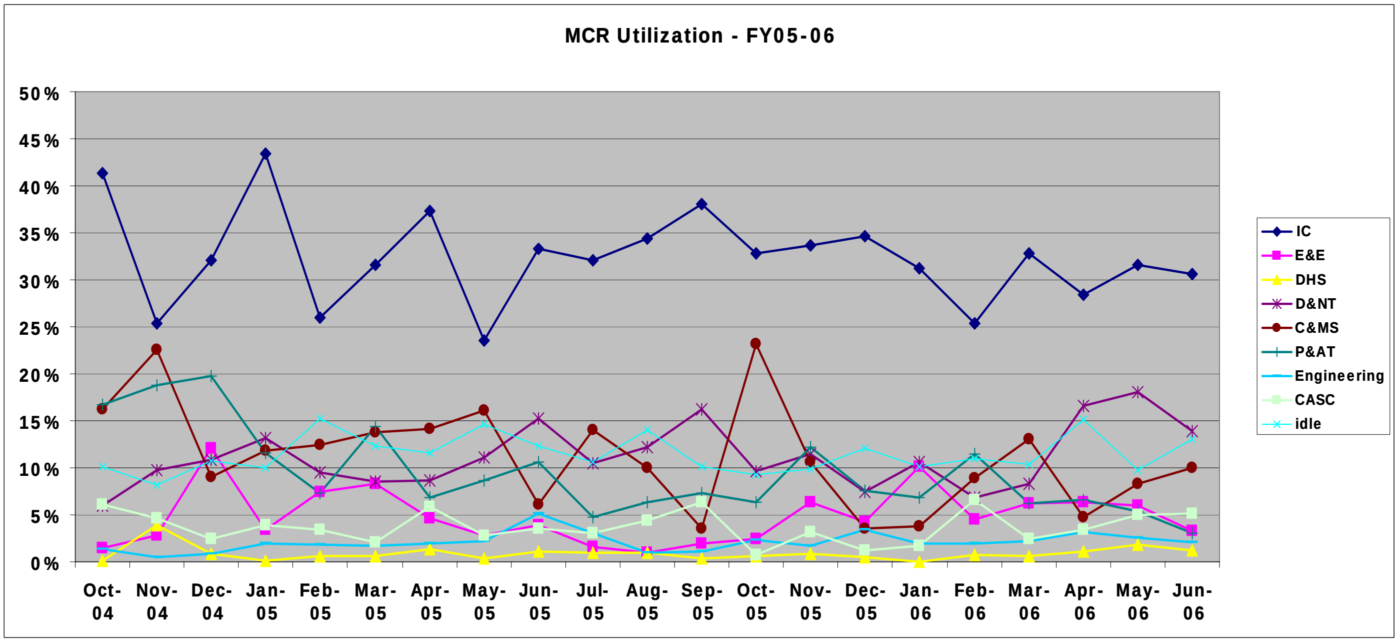


Figure 1.3.1 Institutional utilization of MCR

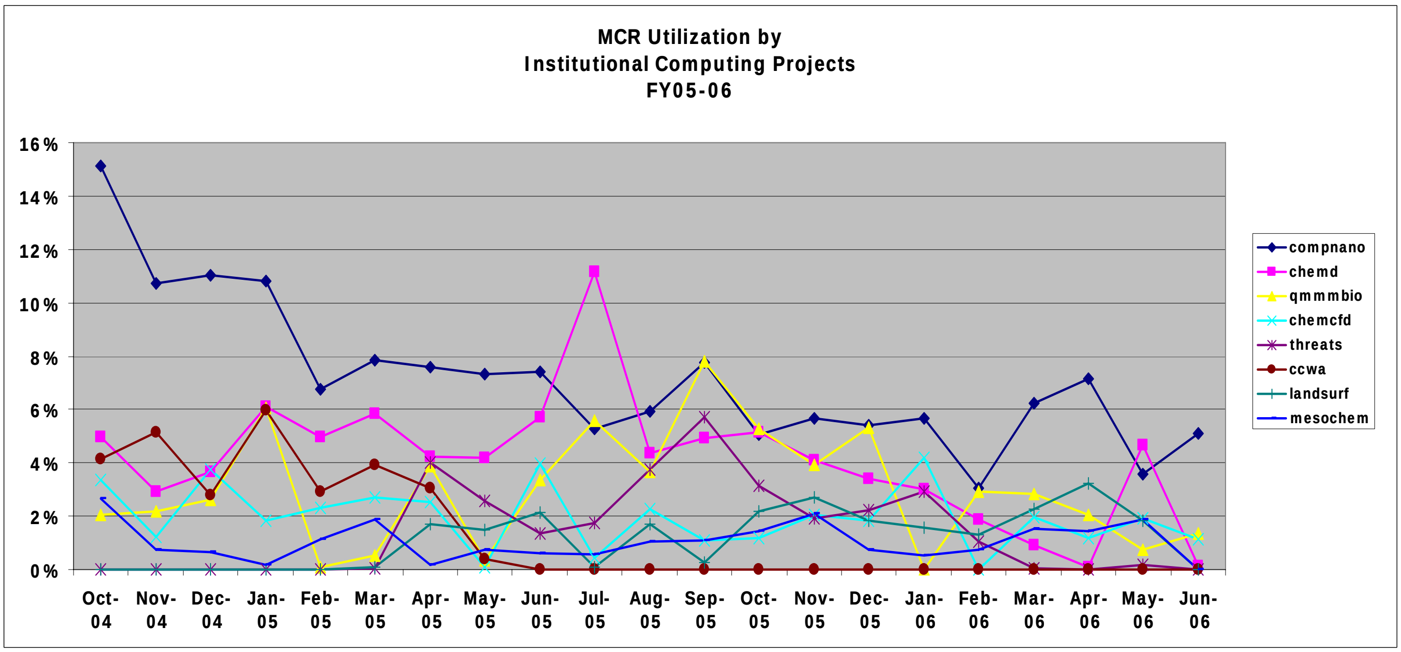


Figure 1.3.2 Institutional utilization of MCR by project name

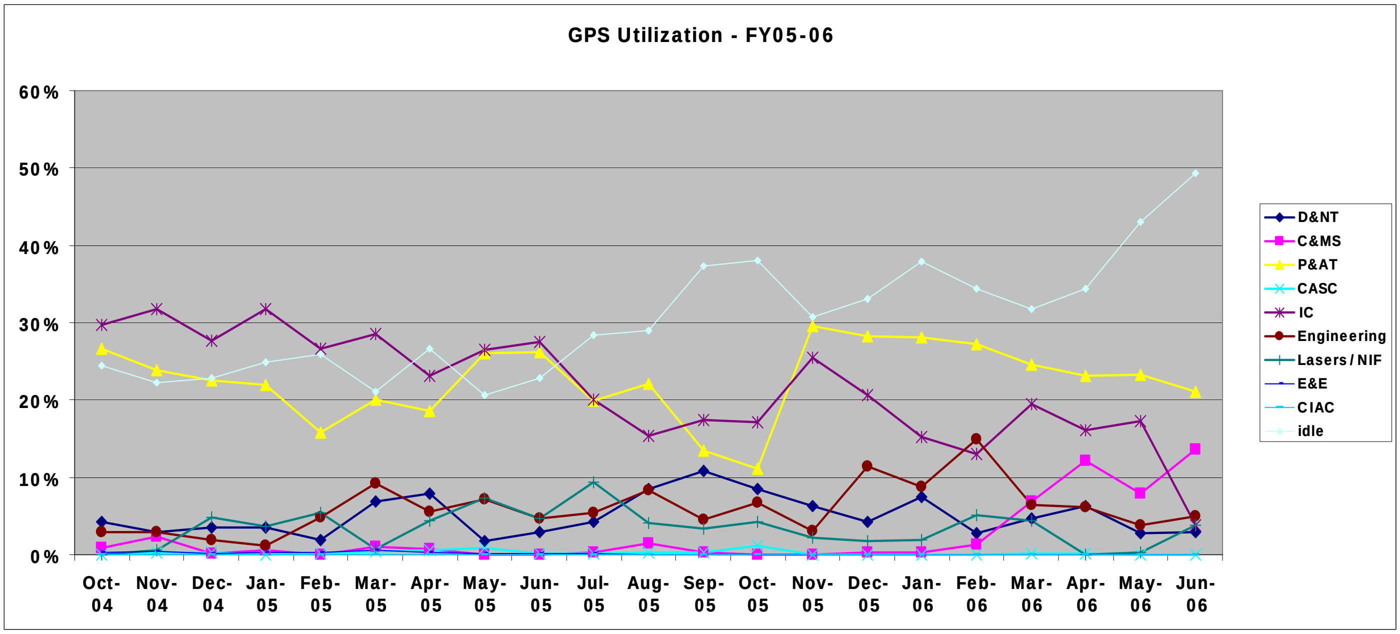


Figure 2.3.3 Institutional utilization of GPS

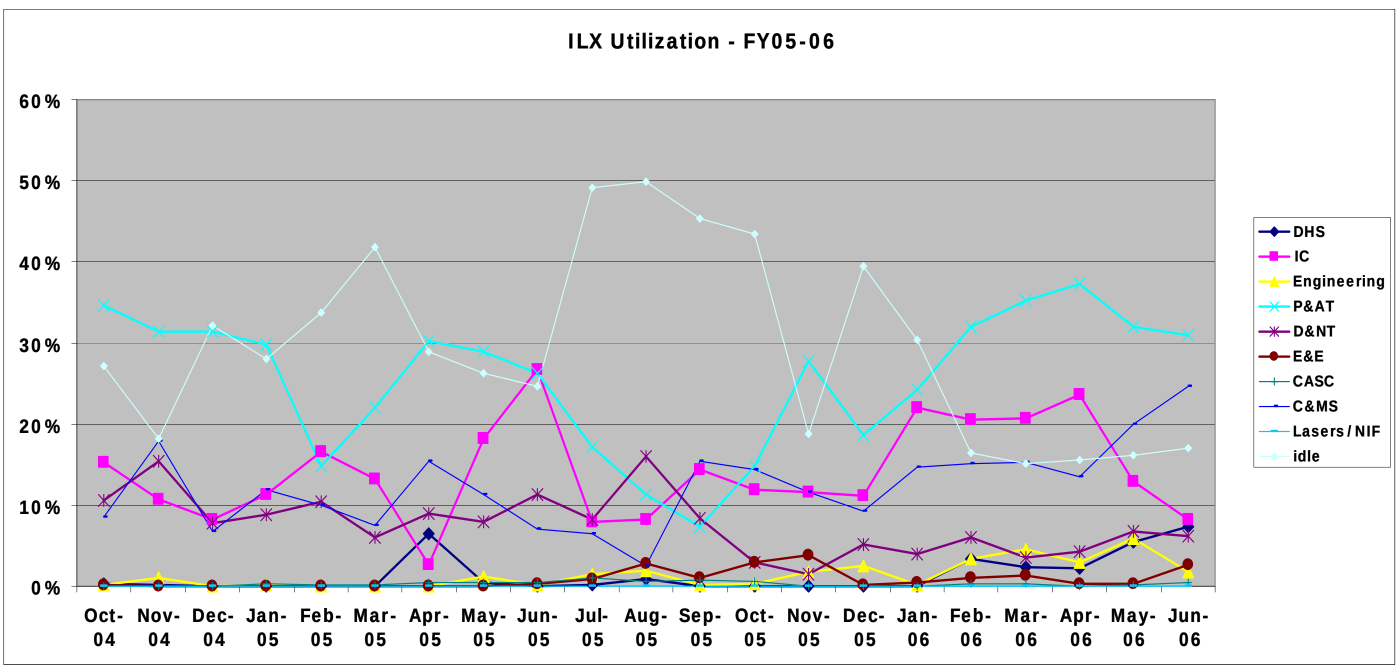


Figure 2.3.4 Institutional utilization of iLX

1.0 MCR and THUNDER Update and Status

MCR and Thunder continue to provide the bulk of compute cycles to M&IC customers. As these systems age and hardware components start to fail, we find ourselves devoting significant resources to keep them operational.

One of our largest challenges this year has been keeping up with hardware repairs on MCR from both the manpower and parts perspective. On December 9, 2005, the hardware maintenance contract for MCR expired. In advance of the maintenance expiration, we purchased a large supply of motherboards, hard disk drives, and power supplies (components with a historically high failure rate). The spare motherboards are the last of this model available for purchase as new parts. In addition to over 100 motherboard replacements, we have replaced well over 200 power supplies on MCR. We had the power supplies made for us as a special order because they were no longer commercially available. Our operating plan throughout the remainder of 2006 is to maintain the node count for as long as possible by replacing parts as they fail. We will do this via our spare parts cache, and when that is exhausted, we will move iLX nodes into the cluster. We have devoted two hardware repair technicians to keeping MCR operational and we still have over 200 nodes out of service waiting for repair. It's apparent that MCR has reached the end of its useful life, and thus it will be retired at the end of the calendar year.

Thunder continues to struggle with compiler, scaling, and file system issues. In spite of all the challenges we have faced, however, M&IC scientists have been able to accomplish some outstanding science (see Section 7). We plan to convert Thunder from the M&IC Capability resource to a Capacity resource when Atlas comes on-line. Thunder is going into the fourth year of what we hope will be a five-year lifespan. (This is consistent with almost all similar systems; see the history chart, **Table 2.1.2**.)

3.1 Development Environment

The Development Environment on M&IC platforms consists of software resources in combination with applications support expertise provided by ICCD staff and vendor consultants employed by the Center. This past year, we have continued to refine the development environment offered on MCR and Thunder, and we are preparing for the acquisition of the Peloton system, which will be Opteron- and Infiniband-based.

We are continually striving to provide more robust and usable systems and to enhance user satisfaction. Frequent software updates on our Linux systems and hardware failures due to aging resources (such as MCR) have caused some amount of user interruption. We have tried to minimize the impact of these interruptions through close interaction with our user community. From a Development Environment Group (DEG) perspective, we have helped users understand the nature of these interruptions and found solutions to them when the issues fell under our purview. We finalized the contract with Intel Solutions Services (ISS) in 2005. The Intel consultants analyzed the CAM climate modeling code and worked with the compiler team to correct issues with OpenMP so that the code could run in a threaded mode. While a consulting contract is no longer in place, we continue to maintain a strong working relationship with Intel and they are responsive to our requests and needs. Intel continues to address compiler issues raised in our error reports. Per our prior request, they have also incorporated LLNL-provided codes into their compiler regression suite, which has helped reduce Intel compiler regressions. To help mitigate the impact of the ISS contract termination, we began a local performance tuning effort. This analysis and tuning (work performed by Intel consultants in our previous contract) was performed by DEG, thereby growing in-house expertise. Several successes were achieved in these efforts. For example, the code that we used to initiate the in-house tuning project (Rocflu) sped up by a factor of 2.5 with our suggested modifications. Speedup of James Vary's MFDN code was large for at least one case. More tuning projects and follow-up work are anticipated for the coming year. We had additional success in analyzing and improving the performance of the system MPI libraries on both Thunder and MCR. Specifically, we identified collective algorithms that were performing sub-optimally. We then recommended and assisted in the development of solutions to improve their scalability. Particular collectives such as "Gather" and "All-to-All" now complete in less than one-tenth of their previous time. This directly improved the performance of several user codes. For example, CPMD now runs twice as fast with these new libraries on Thunder.

As MCR ages, we are focusing less on adding new software and more on upkeep of our existing offerings. We did, however, add the PathScale compilers in the past year. This was partly a method for testing out these compilers before the arrival of Peloton, on which they will be a premiere compiler. The Intel compilers, still superior in performance to the GNU and PGI compilers for most applications, continue to be our compilers of choice. The PathScale compilers, however, are superior for select applications, and while we are not recommending them over the Intel compilers (because the benefit is not universal) they are proving to be a suitable compiler alternative. We also added the FlexeLint code correctness tool to all platforms at the request of an M&IC user. Thunder continues to have fewer tool offerings than MCR, and we continue to rely on the Intel compilers to achieve optimal performance on IA64. Some challenges arose with the release of the Chaos 3 Linux OS, but these were related to hardware and some atypical software bugs. Proper release planning and vendor interaction prevented software incompatibility issues as happened with the Chaos 2 release.

A large number of available tools are open source, which increases local support requirements. See **Table 3.1.1** for some highlights of available tools.

Tool	Vendor	Current state of releases
Intel compilers (support OpenMP)	Intel	Production releases for: IA32: 8.1, 9.0, 9.1; IA64: 8.1, 9.0, 9.1; x86-64 (Opteron, em64t): 9.0, 9.1
PGI compilers (support OpenMP)	PGI	Production releases for: IA32: 6.0, 6.1 x86-64 (Opteron, em64t): 6.1
PathScale compilers (support OpenMP)	PathScale	Production release version 2.1 for IA32 and x86-64 (Opteron, em64t) only.
GNU compilers	Open source	Various production releases available, 3.4.4 is the current default.
Quadrics MPI	Quadrics (open source)	Production release is based on MPICH 1.24. Current version for both Elan 3 (IA32) and Elan 4 (IA64) is: qsnetmpi 1.24-48.intel81
OpenIB MPI (MVAPICH)	Ohio State University (open source)	For x86-64 Infiniband systems, MVAPICH 0.9-7 is current; MVAPICH is based on MPICH and is layered on top of OpenIB stack.
TotalView	Etnus	Production and beta releases available (7.0.1-5-LLNL is default -- includes LLNL-specific mods).
PAPI (hardware counter tool)	U of Tennessee (open source)	Available on all platforms.
Valgrind (memory correctness tool)	Open source	IA32: Version 3.0.1 IA64: Not available x86-64: Version 3.2.0
Vampir, Vampirtrace (Parallel code profiling tool—called Intel Trace Analyzer and Collector)	Intel, Pallas	IA32: 5.0.0 IA64: No longer supported on IA64 x86-64: Not available Vampir has been deprecated in favor of VNG (see below)

Tool	Vendor	Current state of releases
Vampir NG (Vampir Next Generation)	TU Dresden	IA32: Version 1.4.0 x86-64: Not yet available
MKL (Math Kernel Library)	Intel	Production and beta releases available on all platforms
AMD ACML (AMD Core Math Library)	AMD	Available for x86-64.
Flint (Fortran Lint)	Cleanscape	IA32 only: Version 5.0.
FlexeLint	Gimpel Software	Version 8.00s available on all platforms
mpiP (MPI profiling)	LLNL	Maintained in-house; production releases available on all platforms.

Table 3.1.1 Tools available on MCR, THUNDER, and current Livermore Computing x86-64 systems (should predict tool suite on upcoming Peloton system)

3.2 M&IC System Schedule

Table 3.2.1 contains the schedule for retirement of existing systems and the projected availability dates for the new Peloton systems.

Task Description	Dates
Zeus Limited availability	09/18/06
Zeus General availability	10/30/06
New Serial capacity Limited availability	10/30/06
New Serial capacity General availability	11/10/06
GPS Retirement	12/14/06
iLX Retirement	01/09/07
Prism Limited Availability	10/06/06
MCR Retirement	01/09/07
Atlas Restricted availability (science runs)	10/16/06
Atlas Limited availability	12/11/06
Atlas General availability	01/08/07
Thunder conversion to capacity	01/08/07

Table 3.2.1 M&IC system schedule

Zeus (11 TERAFLOPS peak) will be the Capacity replacement for MCR. Atlas (44 TERAFLOPS peak) will become the new M&IC Capability resource replacing Thunder. Prism will be a new visualization cluster for the unclassified systems. Thunder will be converted to capacity to help Zeus meet the demand for parallel capacity. GPS and iLX will be replaced by the new serial capacity nodes. All new parallel and serial M&IC systems are based on the same node architecture. The slides below provide some of the details of the Peloton procurement and the winning architecture:

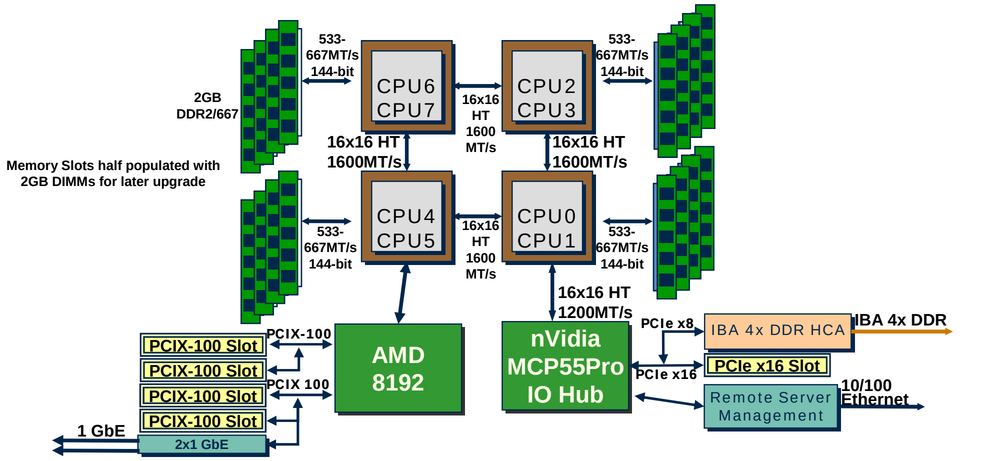
M&IC Linux Capability Cluster (Peloton) Procurement

- Procurement Strategy
  - Define consistent ~5 TF/s SU with room for upgrades
    - 2-socket dual core Xeon (4FP/clock) or 4-socket dual core Opteron
    - IBA 4xDDR with improved Mellanox HCA or PathScale IBA 4xSDR
    - Standard SW for Build and Acceptance: Chaos on RHEL V4, Lustre, OpenIB, MPICH2, SLURM/LCRM, Synthetic Workload (SWL) testing
  - Options for multiple clusters
    - Upgrade 1 SU development cluster to 2SU config (minimal cost)
    - Additional 4SU for classified White replacement
- Enable fast path from build to production
  - Standard SU (HW+SW) for reproducibility
  - Vendor builds SU with SWL pre-ship
  - Vendor delivers SU to site with SWL acceptance
  - Vendor aggregates SU into cluster with SWL acceptance

Peloton Selected Appro

- M&IC Linux Capability Cluster
  - Results from Peloton procurement action
    - 12 bidders and 14 bids
    - Aggressive designs with six in competitive range
    - Award to Appro based on best value to the University
  - M&IC LTO includes 55.4 teraFLOP/s at $13.9M
    - 8 SU 44.3 teraFLOP/s Atlas capability cluster
    - 2 SU 11.1 teraFLOP/s Zeus capacity cluster
- Appro Strengths
  - Superior technical solution
  - Excellent price
  - Long term partnership in Open Source development
  - Small company located in Fremont with $6B/yr Synnex doing contract manufacturing and supply chain management
  - Excellent track record of over achievement

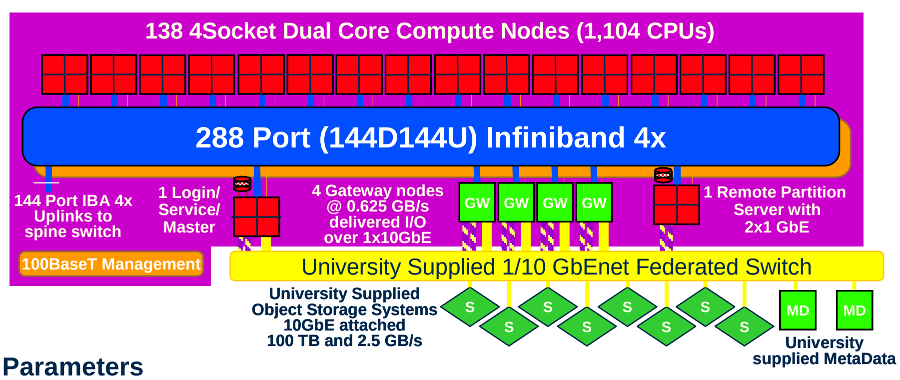
SuperMicro H8QM8-2 Node Block Diagram Dual Core Socket F Opteron



Dual core Socket F Opteron at 2.4 GHz (95W) and 2 GB DDR2/667DIMM
- **Node peak is 38.4 GF/s**
- **16 GB memory (B:F=0.42) – half populated for later upgrade**
- **42.7 GB/s memory BW (B:F=1.11)**
- **2+2 GB/s IBA 4x DDR BW (B:F=0.10)**
- **Mellanox IBA 4x DDR HCA in PCIe 8x slot**
- **1U form factor**

Upgradeable to Deerhound (4 FP/clock * 4 Core is a 4x boost in peak)

Appro 4xSocket Dual Core Opteron SU System Architecture for 144 nodes, 5.53 TF/s peak

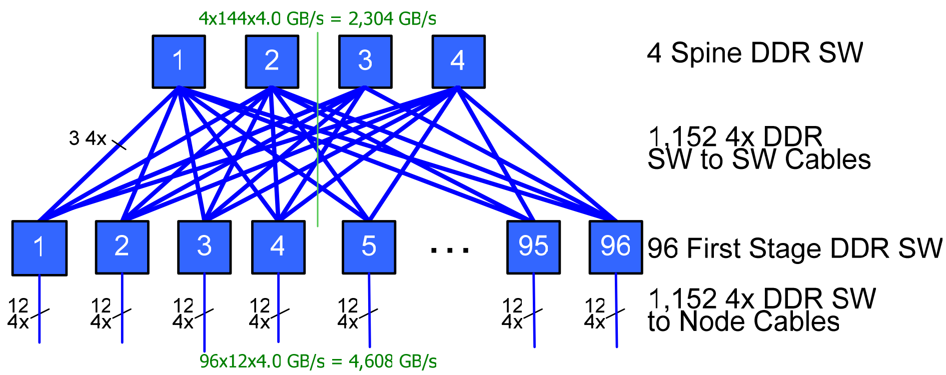


System Parameters
- 38.4 GF/s quad socket F 2.4 GHz dual core AMD (95W) SMP nodes with 16.0 GB, 42.7 GB/s DDR2/667 SDRAM (memory B:F=0.42, BW B:F=1.11)
- <3 μs, 4 GB/s MPI latency and Bandwidth and 8.3M msgs/s over IBA 4x DDR (B:F=0.10)
- Support 800 MB/s transfers to Archive over Jumbo Frame 10Gb-Enet and IBA links from Login node.
- No local disk. Remote boot and SRP target for root and swap partitions on RAID5 device for improved RAS
- IO Bandwidth 2.5 GB/s (B:F=0.00005) delivered parallel I/O performance
- Disk Capacity 100 TB (B:F=18) global parallel file system in multiple RAID5

Note: Socket F can be later upgraded to Deerhound

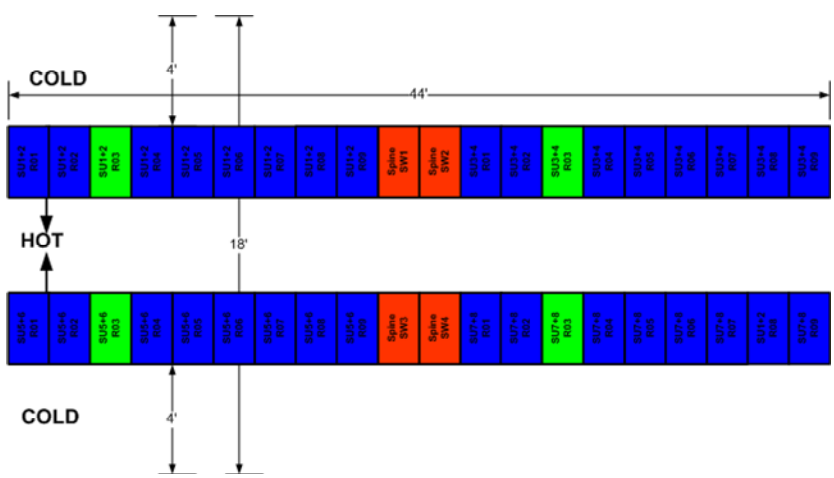
IBA 4x DDR fat-tree interconnect for a 8xSU 1,152 node, 44.3 TF/s Atlas cluster

Optimized design with 24P switches
- **2.5x cheaper**
- **Improves SU density**
- **Reduces first stage cable length**
- **Does not impact MTBF**



Atlas Capability Cluster is an extremely dense solution

Atlas cluster is 1,152 nodes with 4,608 sockets and 9,216 cores with a peak of 44 teraFLOP/s and 18.4 TiB of memory. That is 3.36x White in 28% of the floor space and 90% of the power/cooling. This machine is more powerful than ASC RedStorm while 5.76x cheaper and would be 6th on the current TOP500 list.

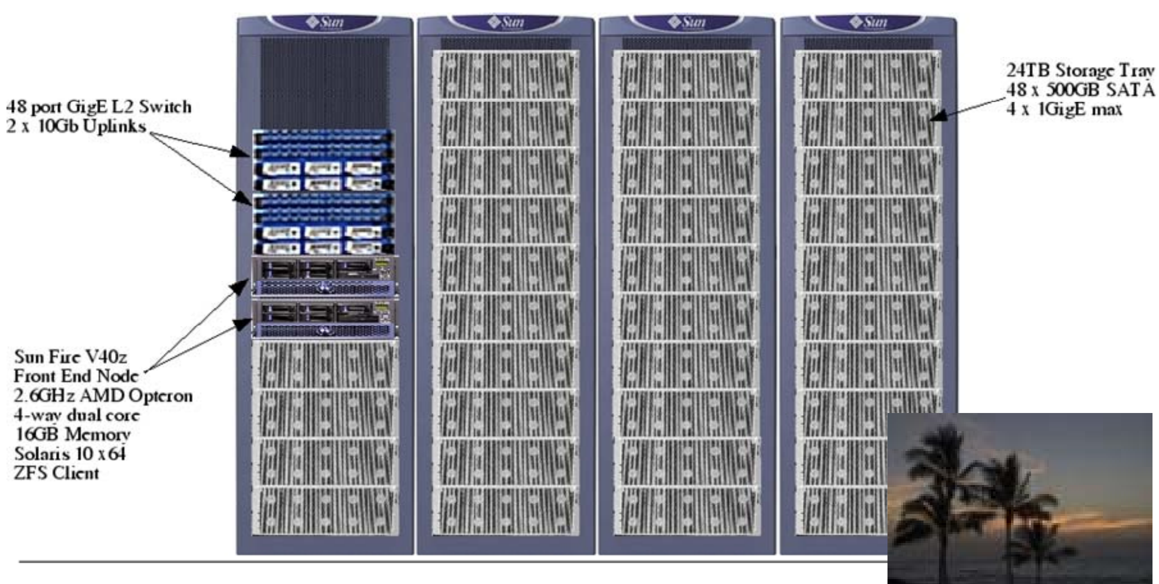


4.0 Green Data Oasis (GDO) Update

The powerful unclassified systems at LLNL give users the capability of producing vast amount of numerical results. Likewise, experimental facilities can also produce very large data sets. In many cases, program goals and science needs require that external partners access these datasets. Many of these external partners are universities that include U.S. citizens, foreign nationals, and sensitive country nationals. Current security rules greatly limit access by these external partners; this is especially true for sensitive country foreign nationals and foreign nationals. Yet international partnerships are necessary to achieve programmatic goals and to showcase LLNL as a world-class science organization. One possible solution to this problem is to create a storage capability outside the LLNL firewall. We are calling this the Green Data Oasis (GDO) Collaboration. Programs across LLNL needing such a capability include PAT, C&MS, NIF, E&E, Biosciences, and Computations. Simulation results and experimental data sets can be placed on that storage system for external access. A particular benefit of this facility is that it will enable students and faculty at various UC campuses to access LLNL science.

We are in the early stages of deployment with a few alpha users. The remainder of calendar year 2006 will be a beta phase. There will be a Lab-wide call for proposals in the fall of 2006 to determine phase 1 usage allocations to begin January 2007.

Green Data Oasis Collaboration with Sun will provide institution with external collaboration data sharing and introduce disruptive file system technology



816 TB Raw Storage

5.0 Securing Allocations on M&IC Systems

MCR and Thunder are both programmatic and institutional resources. This means that a science team can gain access to them if a program that has invested in M&IC is willing to provide an allocation to this team based on the program's ownership in the resource. It also means that a team can gain access through a request to the Institution (through a proposal). In this case, the scientist will have an allocation drawing from the Institutional bank. Both processes are described in detail at http://www.llnl.gov/icc/lc/mic/. Select either "Multiprogrammatic Computing" or "Institutional Computing." Here, we provide a synopsis.

5.1 Multiprogrammatic Computing

Co-investing programs will realize allocations in proportion to their investment in MCR. Some of the Institution's allocation will be donated to co-investing programs as a bonus. This bonus is currently 30%. This means that a program investing once at the level of $100,000 will receive $130,000 worth of ownership rights (through the offices of a fair share scheduler). Knowing income in advance is useful to M&IC, so we reciprocate by providing allocations upon receipt of a pledge. The Institution will assume the costs associated with fielding the system, including system administration, power, etc., for at least three years. This further increases the program's leverage.

5.2 Institutional Computing

Any researcher can apply for an allocation by writing a proposal. The process for Laboratory Directed Research and Development (LDRD) researchers is simpler than it is for others, since it is clear that the LDRD PI has already had the work reviewed by the LDRD committee for quality and institutional relevance. For the LDRD researcher, it is merely a question of the magnitude of the allocation. For other researchers, the proposal must also show institutional relevance and computational quality. Web-based forms are available for either case; see http://www.llnl.gov/icc/lc/mic/micdescrp.html for more information.

Each year, if the number of requests is substantial, the ICEG will review the proposals and recommend an allocation using a peer review process. M&IC management will then convene a small group of ICEG representatives to review all allocation recommendations for consistency and availability. M&IC management will then deliver the final ICEG recommendations to the Office of the Deputy Director for Science &Technology (DDS&T) for final review. We have asked that all future LDRD calls include M&IC proposal instructions for justifying allocation requests. Therefore we will no longer be issuing a separate call for institutional cycles as we have in the past. If you need cycles for your institutional project, you must fill out and submit the form from our website. New allocations will be awarded twice per year, in mid-November and mid-May.

This fall, similar to what was done for Thunder, a “grand challenge” call for proposals will be issued by the DDS&T to allocate institutional cycles on Atlas. A small number of proposals (10-12) will be selected to receive very large allocations. To be considered, proposals must address a significant and compelling Grand-Challenge-scale, mission-related problem that shows great promise of achieving unprecedented discoveries in a particular scientific and/or engineering field of research. Project success should result in high-level recognition by the scientific community at large. This will be a separate call from the Green Data Oasis call mentioned earlier.

The ICEG, working with additional reviewers selected by the DDS&T and the directorates, will make recommendations to the DDS&T and Laboratory Science and Technology Office (LSTO) for final decisions. Four criteria will be used to evaluate proposals:

- quality of science and/or engineering
- significance and impact of access to resources
- ability to effectively utilize high-performance, institutional computing infrastructure
- alignment with the Laboratory Science & Technology Long-Range Plan

The DDST and the LSTO will make final Grand Challenge computing allocation decisions. Awards will be announced and user accounts will be established in December. We expect that the machine will be in full service, with all user accounts in force, by January 8, 2007.

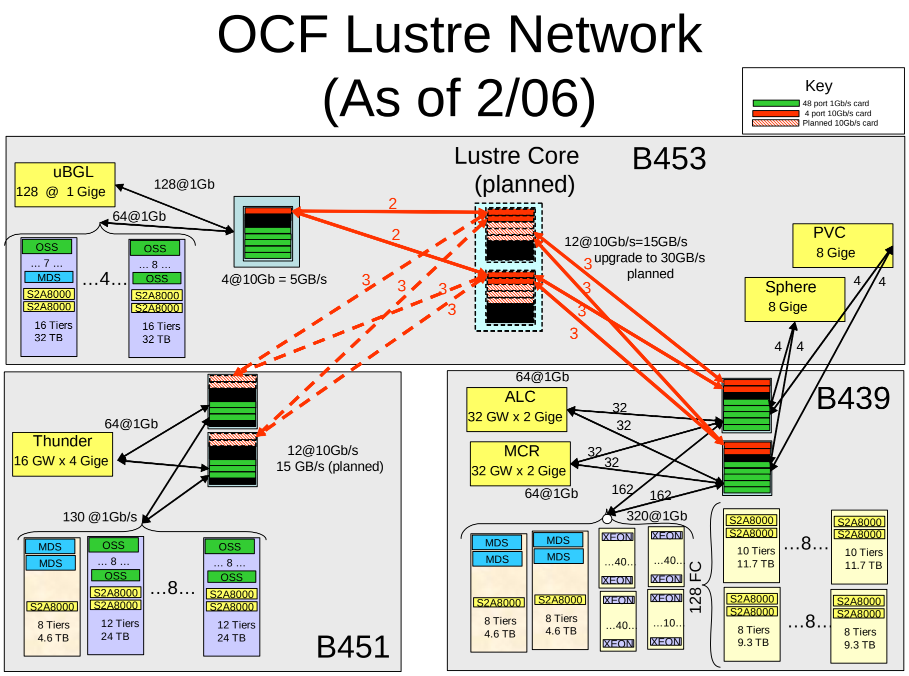
6.0 Lustre File System Upgrade

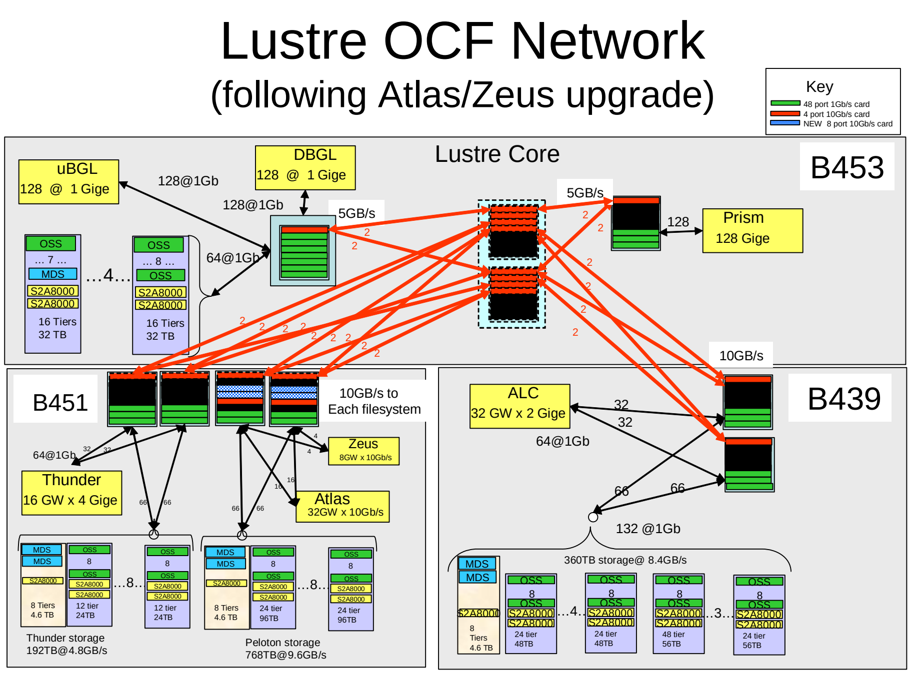
This year most of our FY06 Lustre file system efforts will be driven by the installation of two new major compute platforms, the 11 TERAFLOPS two scalable unit (SU) “Zeus” system and the 44 TERAFLOPS eight SU “Atlas” system. When developing our storage strategy for these new systems, we found ourselves in an interesting position. Our existing network, Lustre object storage servers, and disk controllers are still useful even though the disk controllers are four years old. By leveraging our existing hardware and upgrading only disks and enclosures, we will be able to provide sufficient bandwidth and a greatly enhanced storage capacity at a substantial cost savings. Unfortunately, we must disrupt current service in order to relocate, upgrade, and reconfigure the storage hardware. As a result of this upgrade, the current MCR and ALC file systems (/p/ga1, /p/ga2, /p/gm1, and /p/gm2) will be phased out. The old file systems will be replaced by two new (and much larger) file systems that will be called /p/lscratch1 and /p/lscratch2. We hope to have both of these new file systems along with Thunder’s filesystem (/p/gt1) mounted on all the major M&IC compute and visualization platforms (Zeus, Atlas, Thunder and Prism) by the end of the calendar year.

The table below shows OCF storage capacity and bandwidth before and after the storage upgrade.

OCF				
	Bandwidth (GB/s)		Capacity (TB)	
	Before	After	Before	After
Thunder (gt1)	5.3	5.3	192	192
uBGL (gbtest)	1.6	1.6	60	60
MCR (gm1)	6.4		93	
MCR (gm2)	3.5		89	
ALC (ga1)	3.2		74	
lscratch1 (new)		8.4		360
lscratch2 (new)		9.6		768
Totals	20	24.9	508	1380

The following diagrams show the OCF Lustre storage architecture before and after the storage upgrade. Note that the "after" diagram shows that MCR, PVC, and Sphere have been retired. However, they will have access to the /p/lscratch1 filesystem until they are actually retired.





7.0 Thunder Grand Challenge Results

With the integration of Thunder, the M&IC program offered LLNL scientists and collaborators access to an unparalleled set of resources for simulation science. No place else in the world offers anything close to this level of High Performance Computing (HPC) capability, bolstered by experienced computer experts dedicated to the enablement of world class science. M&IC science simulation breakthroughs have had major impacts in several areas of research that are important to our national interest. **Appendix A** shows a few result slides from the first set of Thunder Grand Challenge efforts. The Thunder presentations were presented by each of the project PIs to the Deputy Director for Science and Technology and several Associate Directors on May 18, 2006.

8.0 Conclusion

Institutional computing has been an essential component of our S&T investment strategy and has helped us achieve recognition in many scientific and technical forums. Through consistent institutional investments, M&IC has grown into a powerful unclassified computing resource that is being used across the Lab to push the limits of computing and its application to simulation science.

With the addition of Peloton, the Laboratory will significantly increase the broad-based computing resources available to meet the ever-increasing demand for the large scale simulations indispensable to advancing all scientific disciplines. All Lab research efforts are bolstered through the long term development of mission driven scalable applications and platforms. The new systems will soon be fully utilized and will position Livermore to extend the outstanding science and technology breakthroughs the M&IC program has enabled to date.

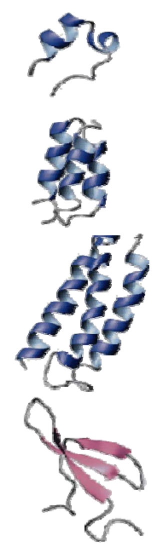
Appendix A.

Farid Abraham, Jed Pitera, William Swope
CAN PROTEINS FOLD AT THE "SPEED LIMIT"?

Downhill Folding Is A New Paradigm

- The generally accepted paradigm of folding is that it is two-state: a free energy barrier to folding exists.
- Recent theory proposed "downhill" folding in the taxonomy of folding landscapes:
  - → folding "speed limit" is achieved when free energy barriers disappear.
  - → proteins as large as 100 residues could possibly be downhill folders and fold in a few microseconds.
- This was an opportunity for simulation to study the 'fast folding' of a protein to completion.

protein	#aa	τ_{fold}
1L2Y	20	4.1
1PRB	47	2.5
2A3D	73	3.0
1E0L	37	30.



Eaton et al, "The protein folding *speed limit*" Current Opinions In Structural Biology 2004,14, 76-88.

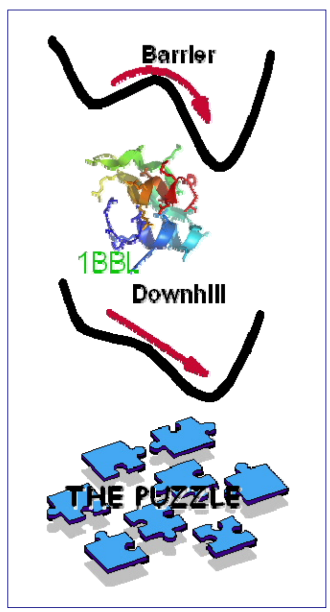
Significant Impact In Protein Folding

New paradigm for folding

- A controversy among experimentalists is resolved by our simulation study on Thunder

General science

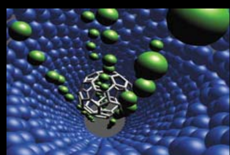
- Demonstrated capability of Thunder scale resources to achieve robust results will help to establish simulation as a growing partner in biophysical research



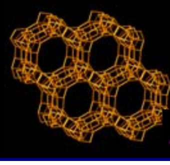
Eric Schwegler - Why study confined water?

The physical properties of confined water play a key role in diverse scientific disciplines and applied technologies

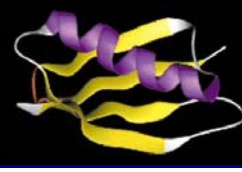
nanofluidics



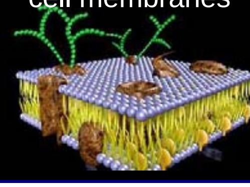
materials science



protein folding



cell membranes



Understanding how the hydrogen bond network of bulk water is modified when water is confined is important for:

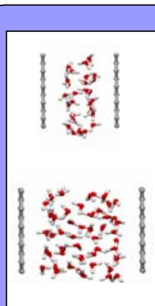
- » Studies of stability and enzymatic activity of proteins
- » Oil recovery
- » Nano-fluidics
- » Heterogeneous catalysis (fundamental role of water-substrate interaction)
- » Corrosion inhibition

Our computational objectives includes specific calculations with available techniques and development of novel simulation tools

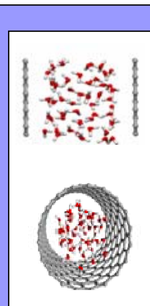
- **Investigate how the solvation of hydrophobic/philic species is affected by confinement**
- **Identify structural "fingerprints" of confined water**
- **Investigate changes in electronic properties**
- **Explore frameworks to define parameters for empirical simulations of confined water, based on *ab initio* results**

Access to Thunder through the Grand Challenge program has enabled us to perform a predictive and systematic study of nanoscale confinement with respect to lengthscales, dimensionality, and interface effects without having to compromise on the level of theory used.

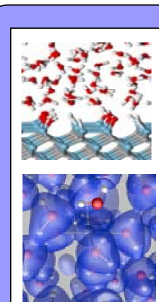
Length scales



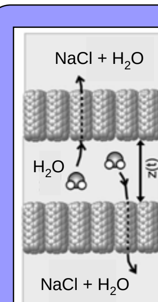
Dimensions



Interfaces/ Solvation



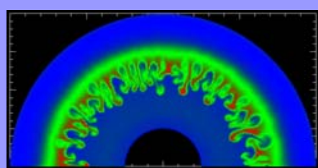
Transport



Vasily Bulatov
Dislocation Dynamics: the promise

Equations governing behavior of individual dislocation are well established and it is possible, in principle, to compute material strength directly by solving these equations

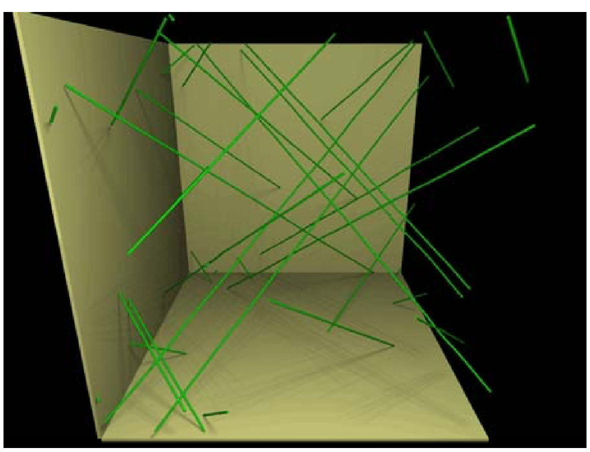
ASC Program at LLNL: material strength under extreme conditions



Key issue – effect of dislocation microstructure on strength

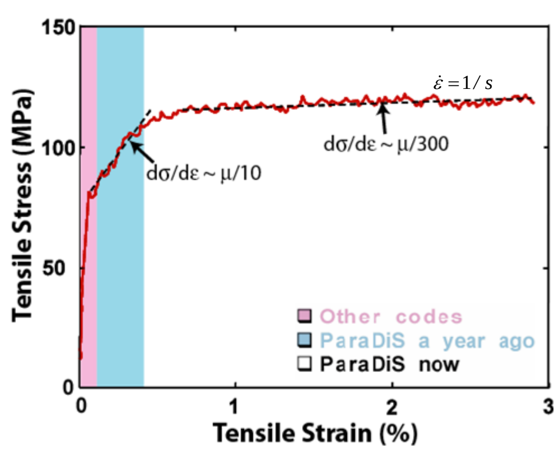
The goal is an accurate, physics-based, experimentally validated and computationally efficient model of crystal strength

ParaDiS meets the challenge



First ever direct calculation of plastic strength of a single crystal across the stages of strain hardening

For the first time, stress-strain behavior can be predicted to large extents of strain with strain hardening



Bulatov et al. *Supercomputing* 2004.

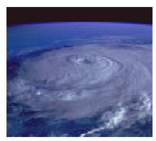
Doug Rotman
LLNL's climate work increasingly examines regional climate science, impacts and adaptation strategies

- Why?
- Because humans and natural ecosystems experience regional, not global, climate
- Because improvements in climate models make meaningful regional projections possible
- Regional climate changes will determine societal impacts and drive climate-related policy decisions

Water Resources

Recreation

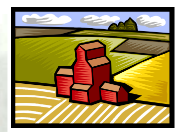
Extreme events



Air quality

Human health

Agriculture

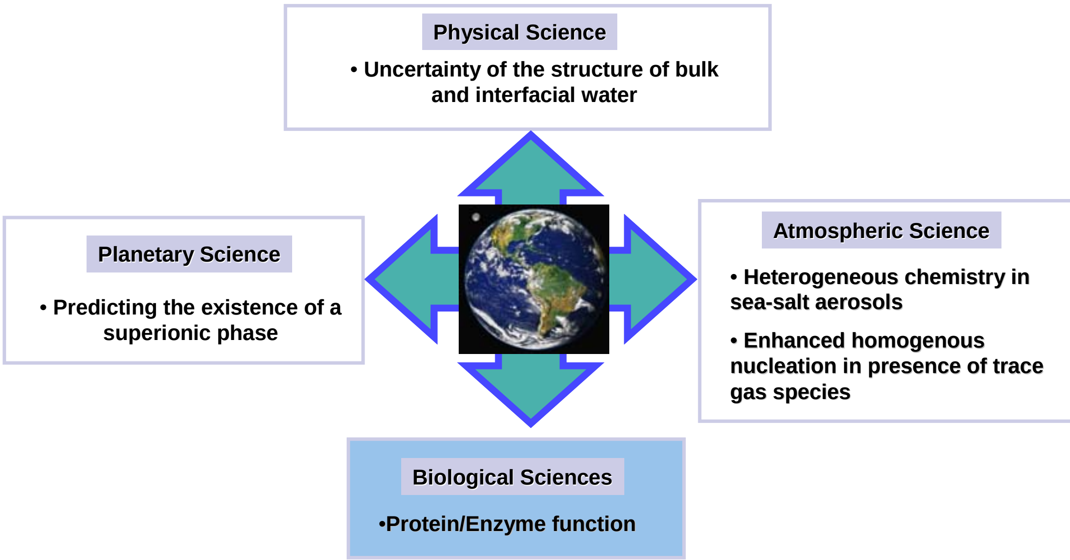


The CCSM3 climate model was run at high resolution for 1100 simulated years on THUNDER

- **Community Climate System Model, Version 3 (CCSM3)**
  - **NSF NCAR climate model, heavily co-funded by DOE**
  - **Basis of DOE's participation in the International Panel on Climate Change**
  - **CCSM contains 5 separate executables running concurrently**
  - **Simulation details**
    - **Atmospheric Dynamical core: Lin-Rood Finite Volume Dynamical Core – *unique***
    - **Atmospheric Physics: standard LW, SW, land, clouds, …**
    - **Ocean: standard LANL POP ocean model**
    - **1 by 1 global resolution, 26 atmospheric vertical layers, 40 ocean vertical levels**
    - **Coupled atmosphere and ocean**
    - **Simulation throughput**
      - **11 years per day on 118 nodes of THUNDER**
      - **We have run 1100 years of simulation**
      - **Our usage has averaged 263 processors of continuous use**
      - **Simulation has been running since last year**

Chris Mundy, Will Kuo, Nir Goldman, Larry Fried

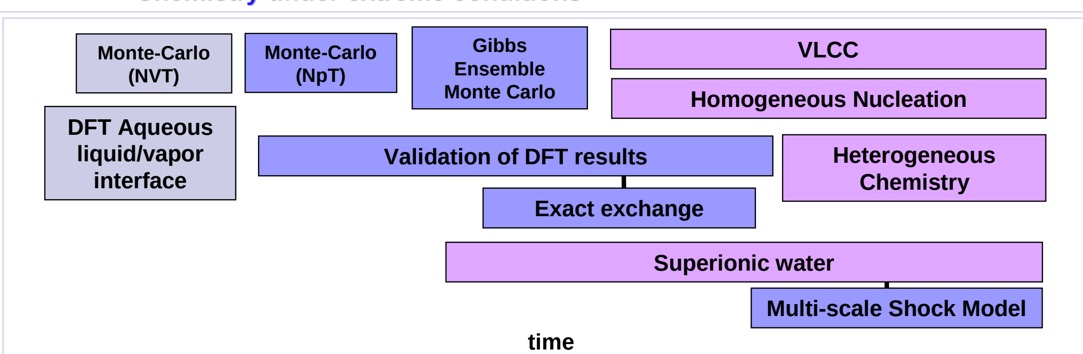
Understanding water and its processes still remains a scientific Grand Challenge for both theory and experiment



Water engenders important phenomena ranging from atmospheric science to biology

Our work on THUNDER answered many unresolved questions about first-principles water in different environments

1. **Prediction of the vapor-liquid coexistence curve (VLCC) for water**
   - **Treating evaporation as a rare event**
2. **Enhancement of homogeneous nucleation of water in the presence of trace gases**
   - **Treating nucleation as a rare event**
3. **The role of heterogeneous chemistry at aqueous interfaces**
   - **System size**
   - **Chemistry as the rare event**
4. **Water in planetary interiors**
   - **Chemistry under extreme conditions**



Our approach to the Grand Challenge has lead to many scientific achievements