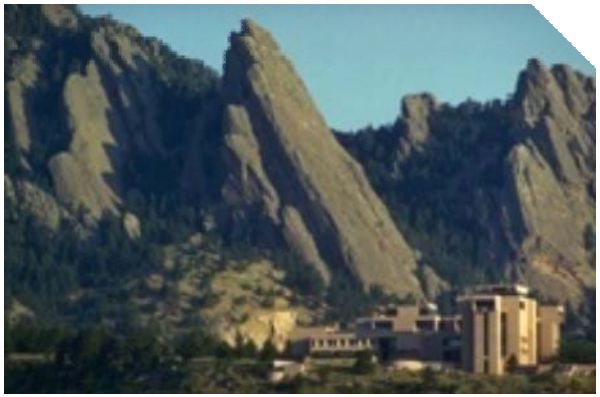
13th ECMWF Workshop on the Use of HPC in Meteorology

Al Kellie
Associate Director, National Center for Atmospheric Research (NCAR)
Director, Computation and Information Systems Lab (CISL)
(Kellie@ucar.edu)

OUTLINE

- A look at how NCAR & CISL are organized
- more of CISL
  - HPC facility
  - Archival storage facility
  - Research data facility
  - Metrics
- Wyoming
- Science efforts

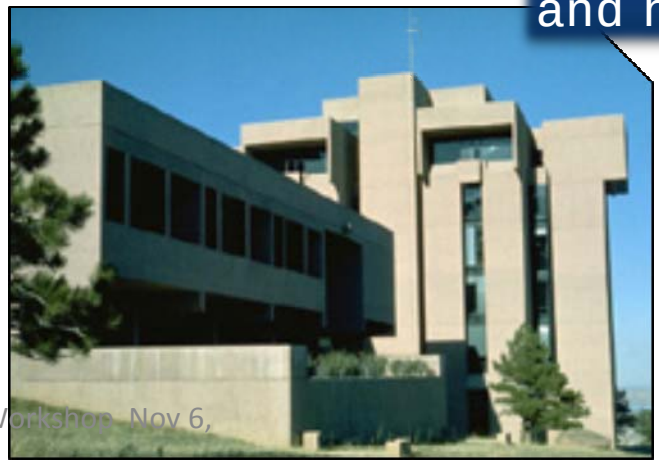


NCAR - a federally funded research and development center sponsored by the National Science Foundation.

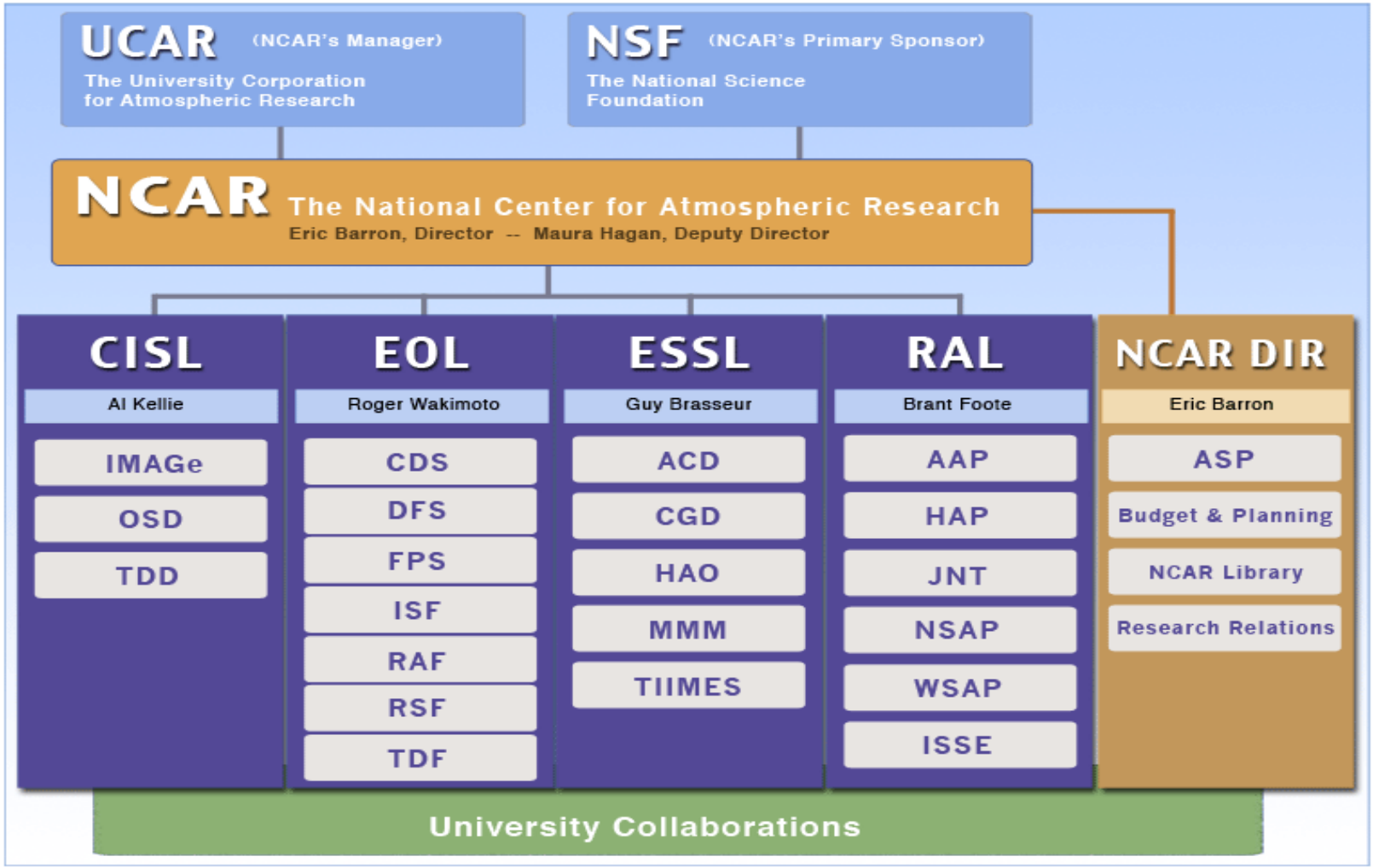
- **Established in 1960 by 14 universities**
- **Managed by the University Corporation for Atmospheric Research (UCAR)**
- **UCAR: non-profit private corporation**
  - **Composed of 73 Member Universities**
  - **18 Academic Affiliate**
  - **46 International Affiliate Institutions**

Principle Objectives:

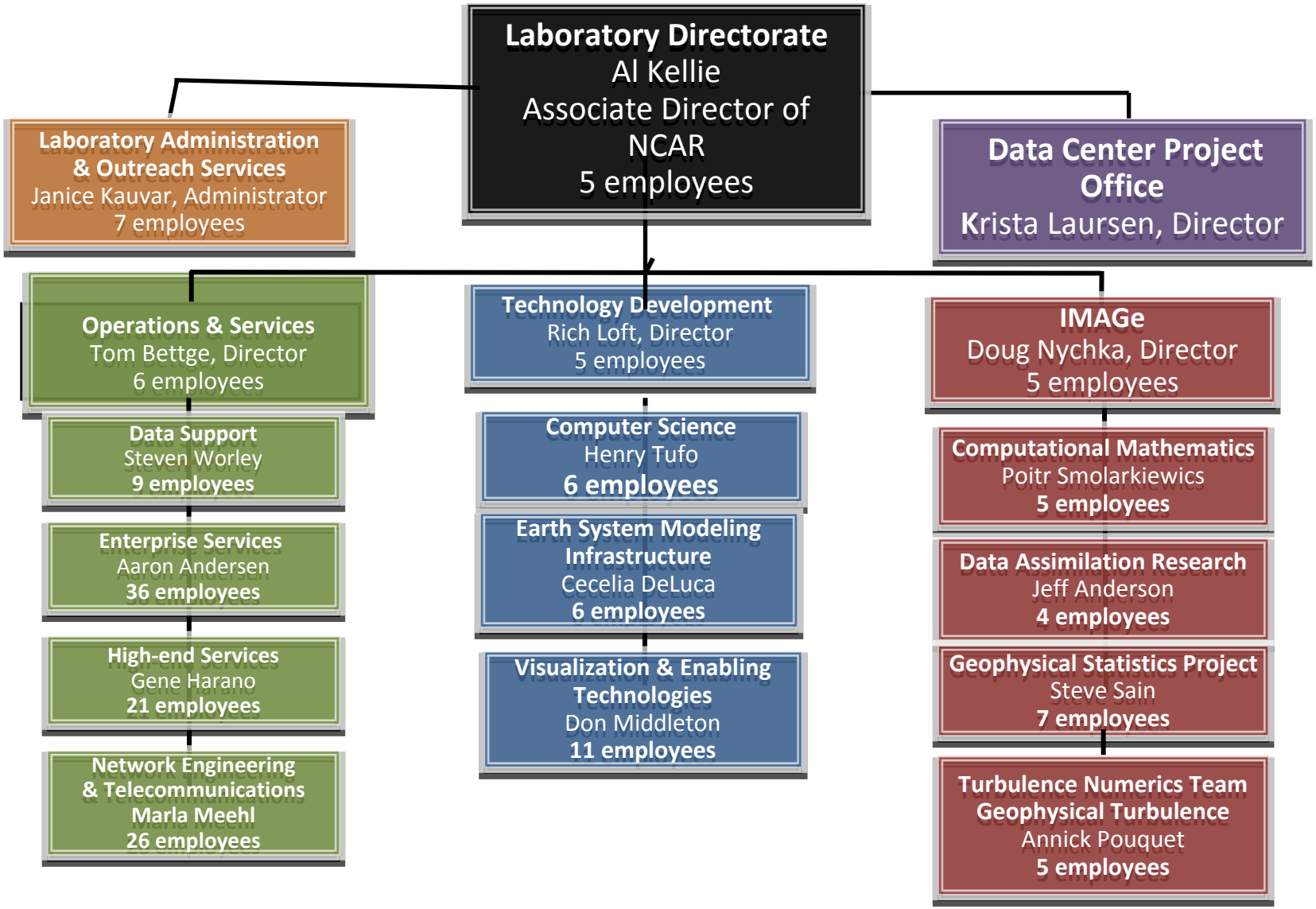
- Partners with universities and research centers,
- Dedicated to exploring and understanding the Earth's atmosphere and its interactions with the Sun, the oceans, the biosphere, and human society.



NCAR Organization



CISL Organization



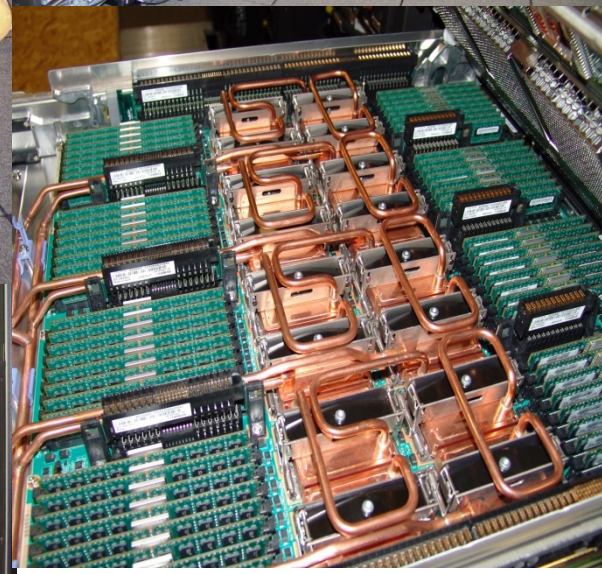
CISL at a GLANCE

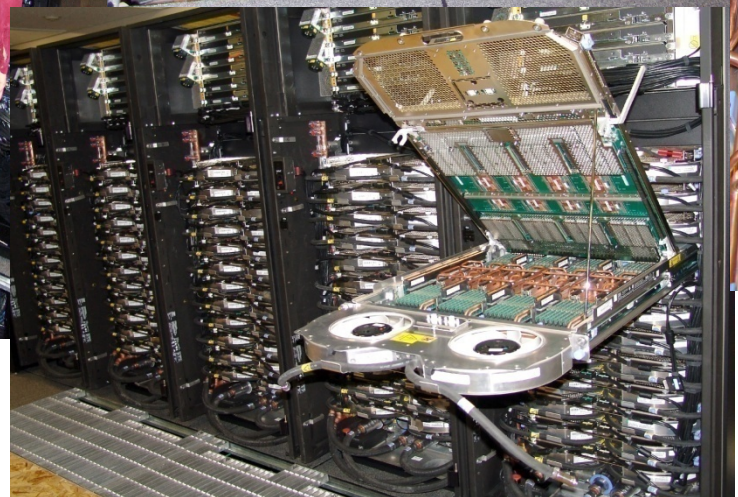
Bluefire (commissioned in June 2008)

- 4,064 IBM Power6 processors, 4.7 GHz, quadrupled NCAR's sustained computing capacity
- 76 teraflops peak
- Hydro-cluster - water-cooled doors and processors 33% more energy efficient than traditional air-cooled, each cabinet weighs 3600 pounds (midsize car)
- 3X more energy efficient than P5+
- Chips run around 140° F compared to 180° F for air-cooled systems
- Runs climate models, atmospheric chemistry, high-resolution forecasts
- LSF job scheduling and queuing system
- 12 TB memory, 150 TB storage
- InfiniBand switch (four QLogic Model 9240 288-port switch chassis)
- Peak bandwidth 6 GB/sec; latency=1.27 microseconds
- 740 kilowatts (60% of our overall computing power)
- Sustained performance: 6-16% of peak for our job mix

Bluefire

- IBM POWER6
- 76.4 TeraFLOPs peak
- Each batch nodes has 32 4.7GHz P6 (dual core chips)
  - 120 batch nodes
  - 69 with 64 GB memory (2 GB/CPU)
  - 48 with 128 GB memory (4Gb/CPU)
  - 2 interactive, 2 share-queue, 4 GPFS and 2 system nodes
- Infiniband switch QLogic 9240 (8 links per node)
- 150 Terabytes disk.
- Sustained Computational Capacity
  - 3.88x that of former P5+
- Computational Capability
  - 1.65x per processor over P5+ for typical NCAR code





ECMWF Workshop Nov 6, 2008

CISL at a GLANCE

Cooling

- Liebert air handlers cool and humidify the air, pulling hot air from the ceiling through a large water-cooled radiator which blows cool air into the raised floor
- 30% relative humidity to reduce static electricity
- Two 450 ton chillers cool the water
- Two 1500 gallon tanks act as thermal sink; store 44° F chilled water; provides 18 min window for chiller failovers (55 seconds without battery)

Power

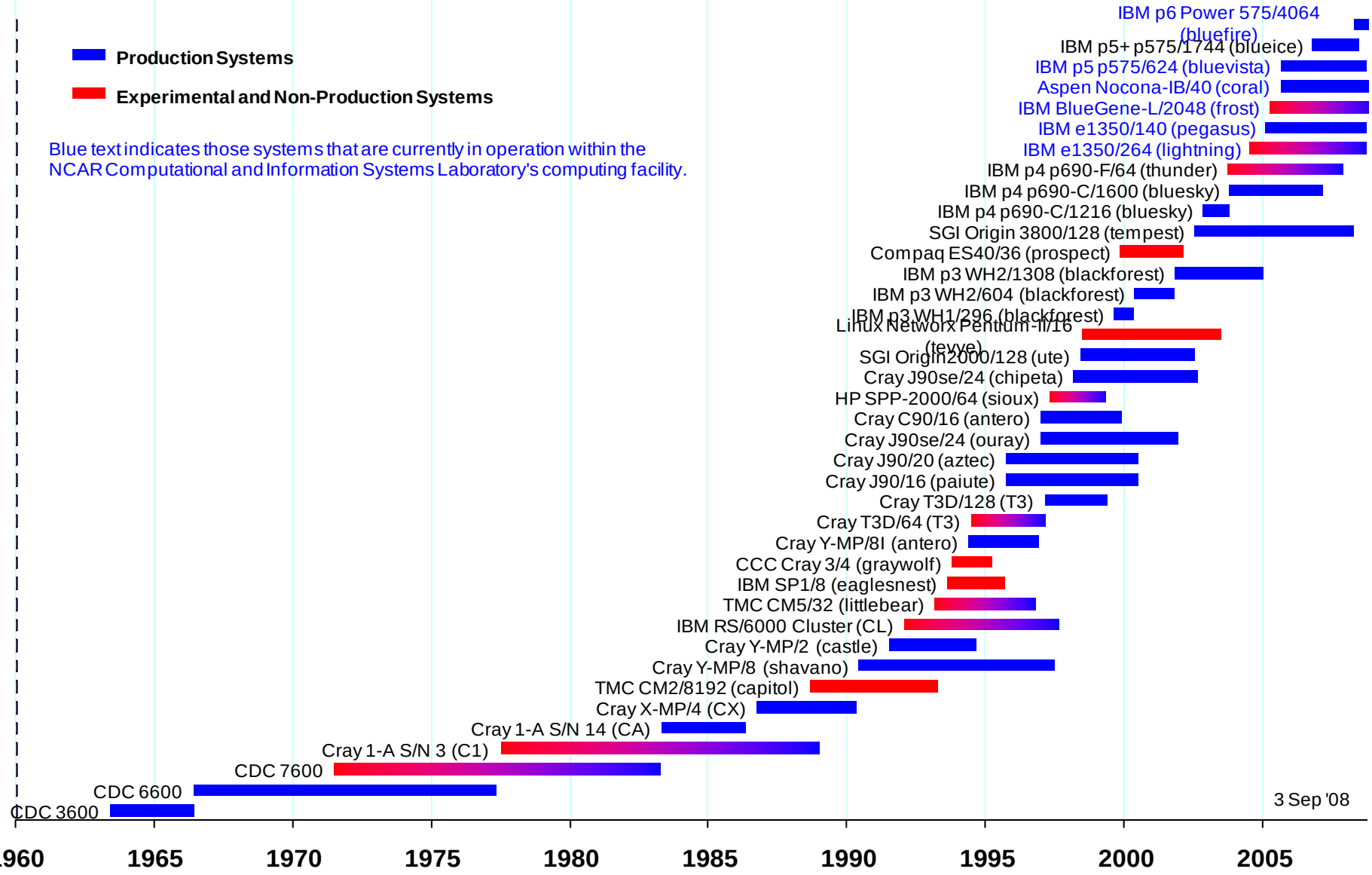
- 2 megawatt facility
- 1.2 megawatts for computing
- 2 Excel feeds of 13,200V each
- $55K monthly power bill
- 60% computing, 40% mechanical
- PowerWare UPS gives us 15 min of 1.2 megawatts
- 2 diesel power generators (1.5 megawatts and 8 hours of diesel fuel each)

CISL at a GLANCE

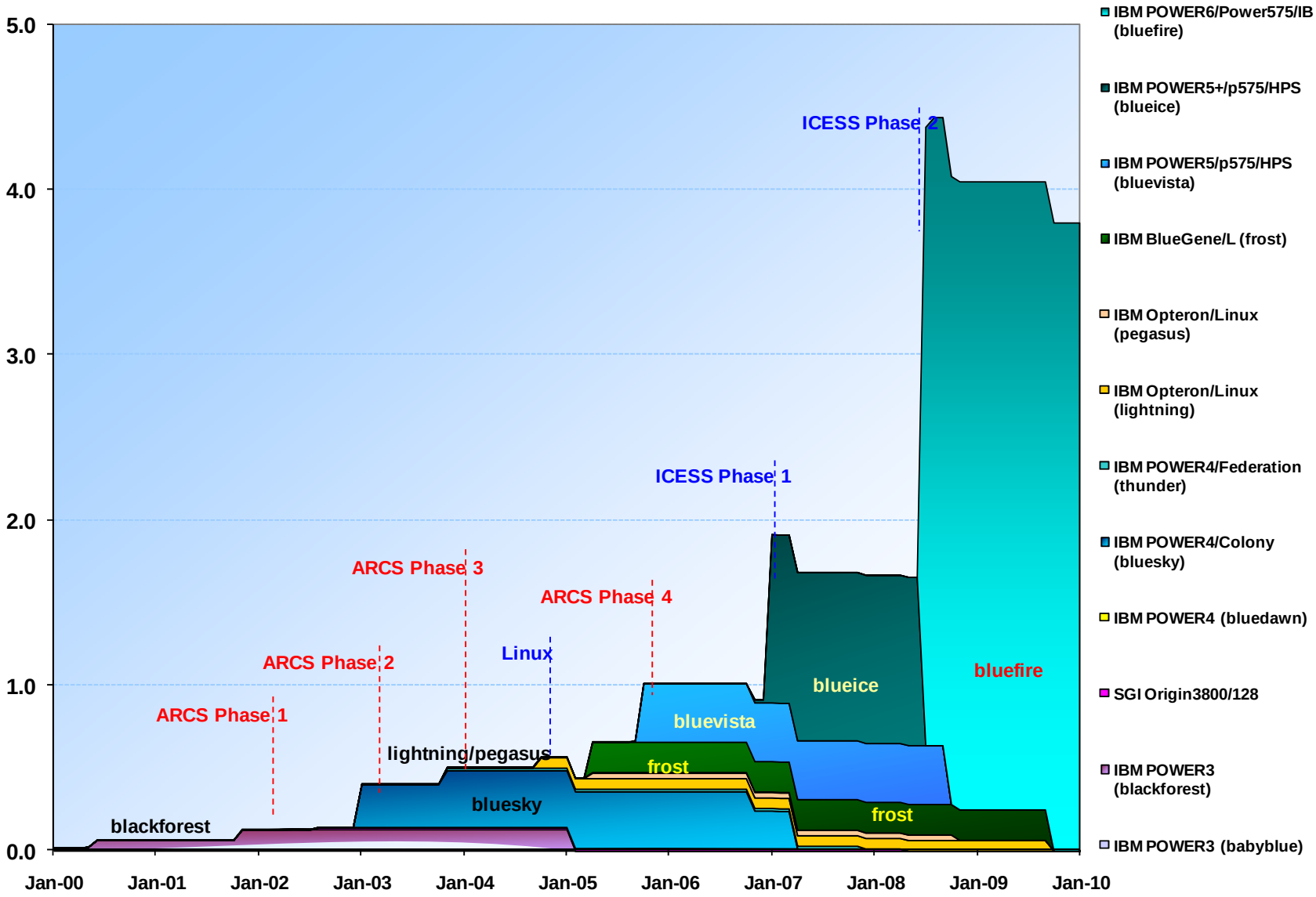
Frost

- IBM BlueGene/L supercomputer
- 2,048 PowerPC 440 processors, 700 Mhz ,5.7 teraflops peak
- Architecture uses densely packed lower speed 700 Mhz processors, with increased bandwidth between processor and memory
- each node in the cluster runs a microkernel rather than a complete operating system
- runs models and code that are optimized for massively parallel computing
- 109 TB storage

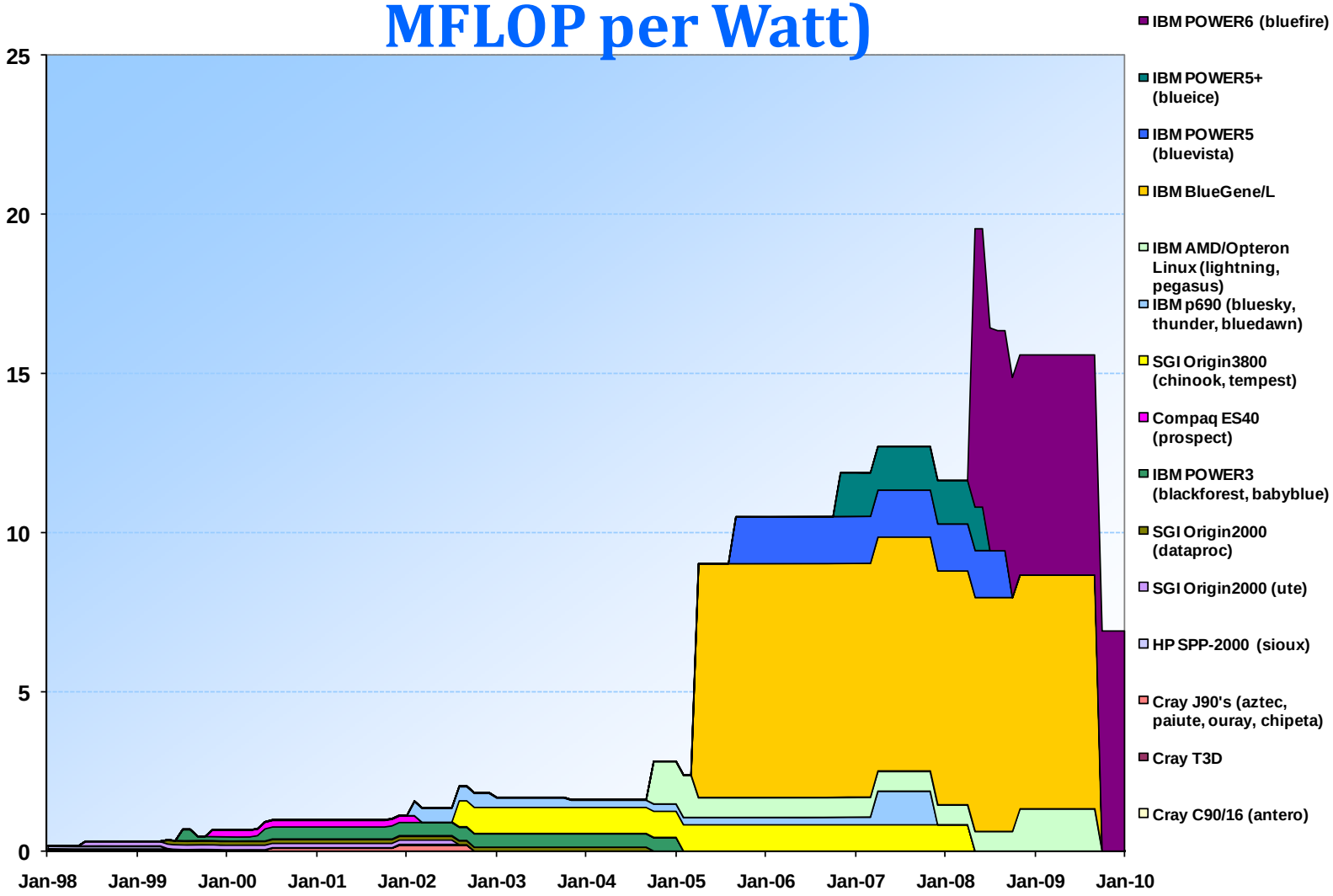
Supercomputing at NCAR



Estimated Sustained TFLOPs at NCAR (All Systems)



Power Consumption (sustained MFLOP per Watt)



CISL at a GLANCE

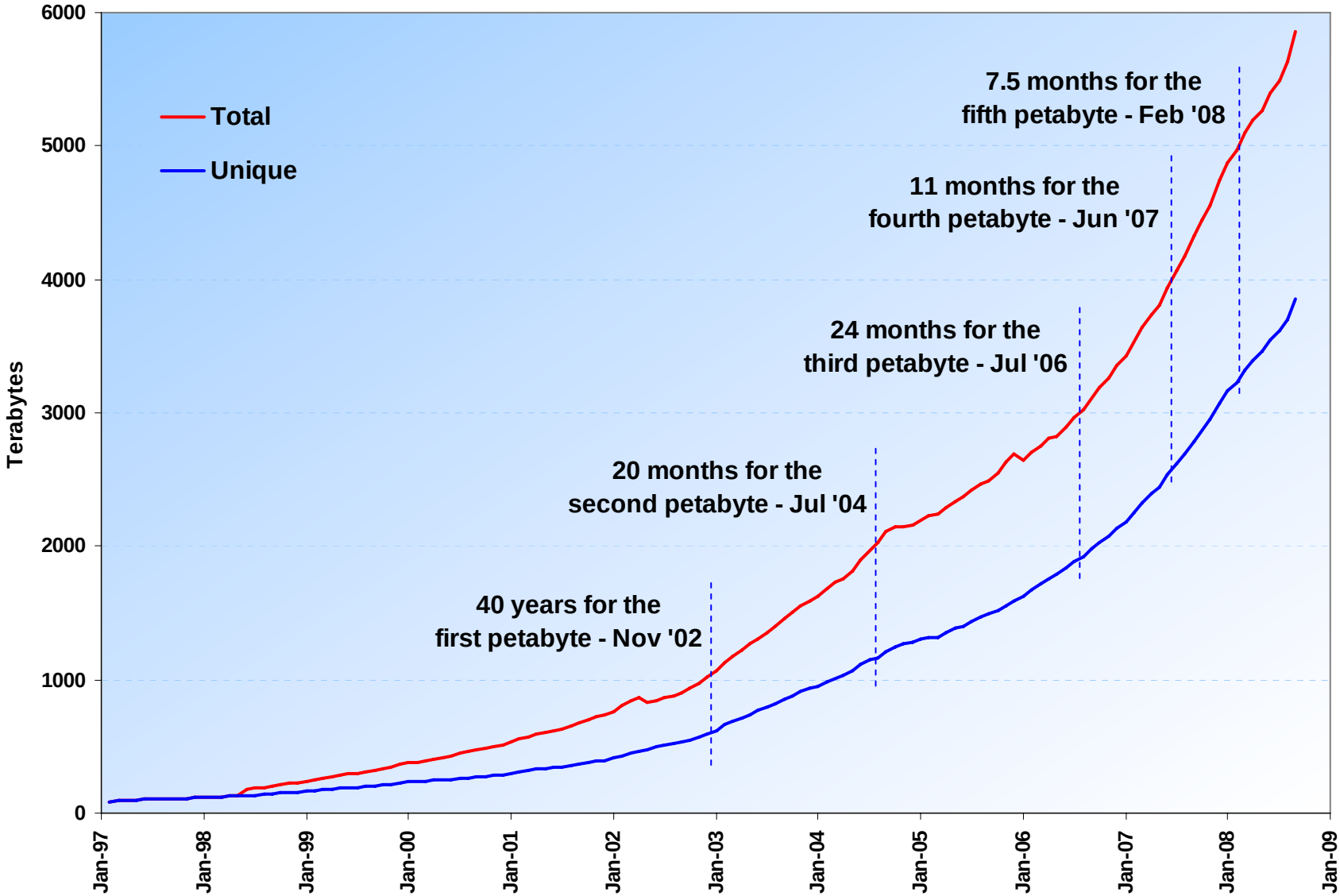
Archival Storage facility (MSS)

- 5 silos, 6,000 slots per silo, 30,000 tapes total
- 200 GB tapes , max capacity of 6 PB has been reached
- Library of Congress print holdings, > 30 million books, were all digitized, it is estimated to be 20 TB (less than 1% of MSS)
- Growth rate increasing with computational rate
- 48 TB disk cache speeds repeated accesses of popular files
- ~ 60% disk cache hit rate for files up to 1 GB
- Massive keeps track of over 50 million files
- MSS software is built in-house at NCAR

Manual Tapes Area

- devices for reading old tapes and media
- tapes found in data warehouses with unique historical data which we read and archive

NCAR MSS - Total Data in Archive



Augmentation of the Mass Storage Tape Archive Resources (AMSTAR)

- Predicted MSS at full capacity by 26 Sept 2008
- Actual, 6PB crossed 27 Sept 2008
- Initiated an procurement for a 4 year contract to augment and/or replace the STK Powderhorn Silos with new robotic tape storage technology, plus developmental HPSS
- AMSTAR Contract signed in early Sept 2008.
- Installations and ATPs underway.

AMSTAR Progression 2008

Phase 1 – **Production Library #1**

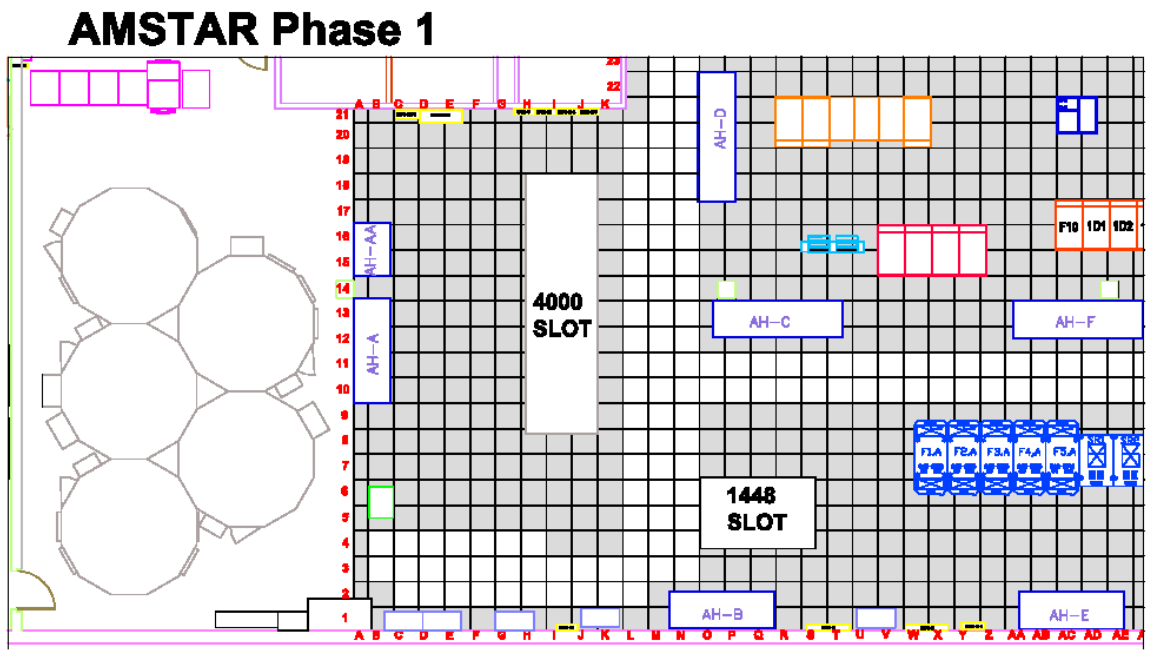
- (1) 4,000-slot SL8500 Library
- (30) T10000B tape drives,
- (4,000) T10000 Tapes,
- (40) T10000 cleaning tapes

•Phase 1a – **Development Library**

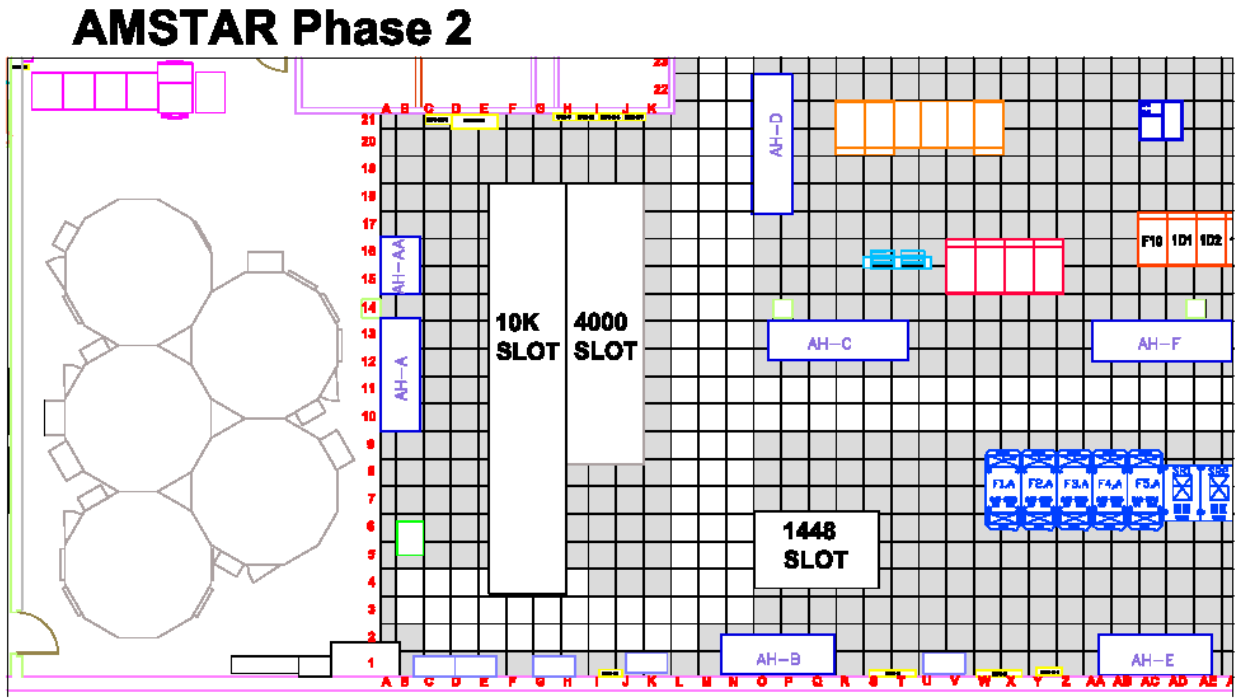
•(1) 1,448-slot SL8500 Library

•(5) T10000B tape drives switch,

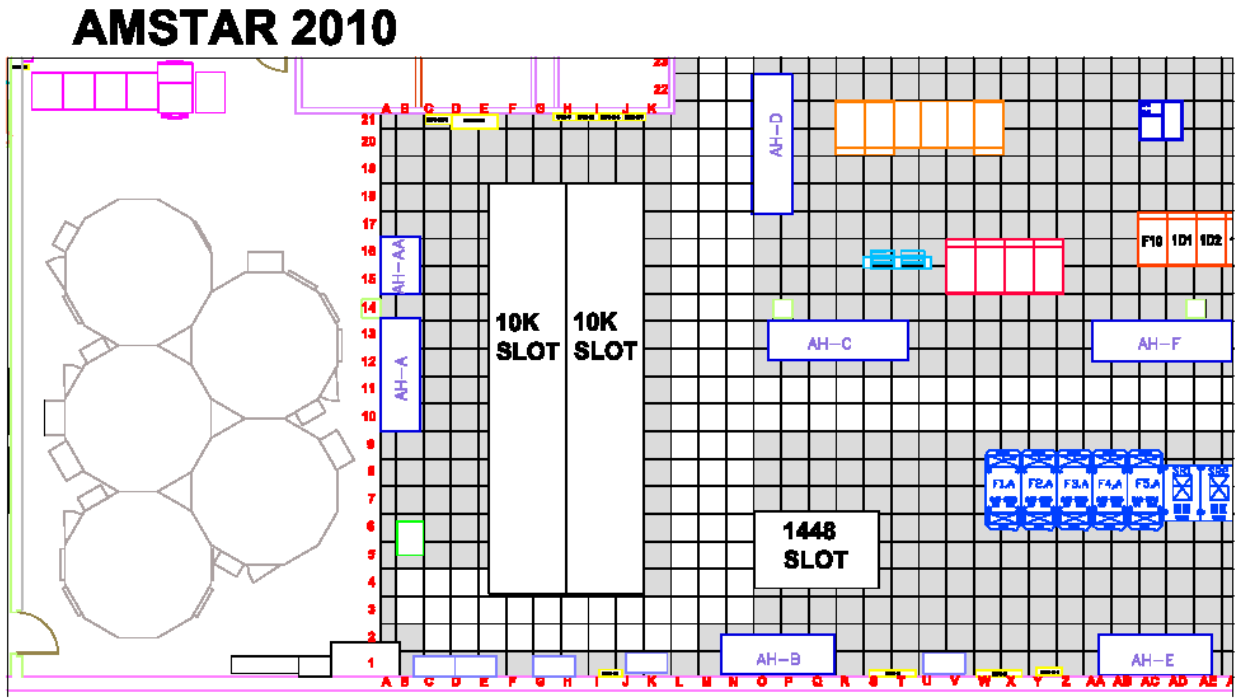
- (1,000) T10000 Tapes,
- (5) T10000 cleaning tapes,



AMSTAR Progression 2009



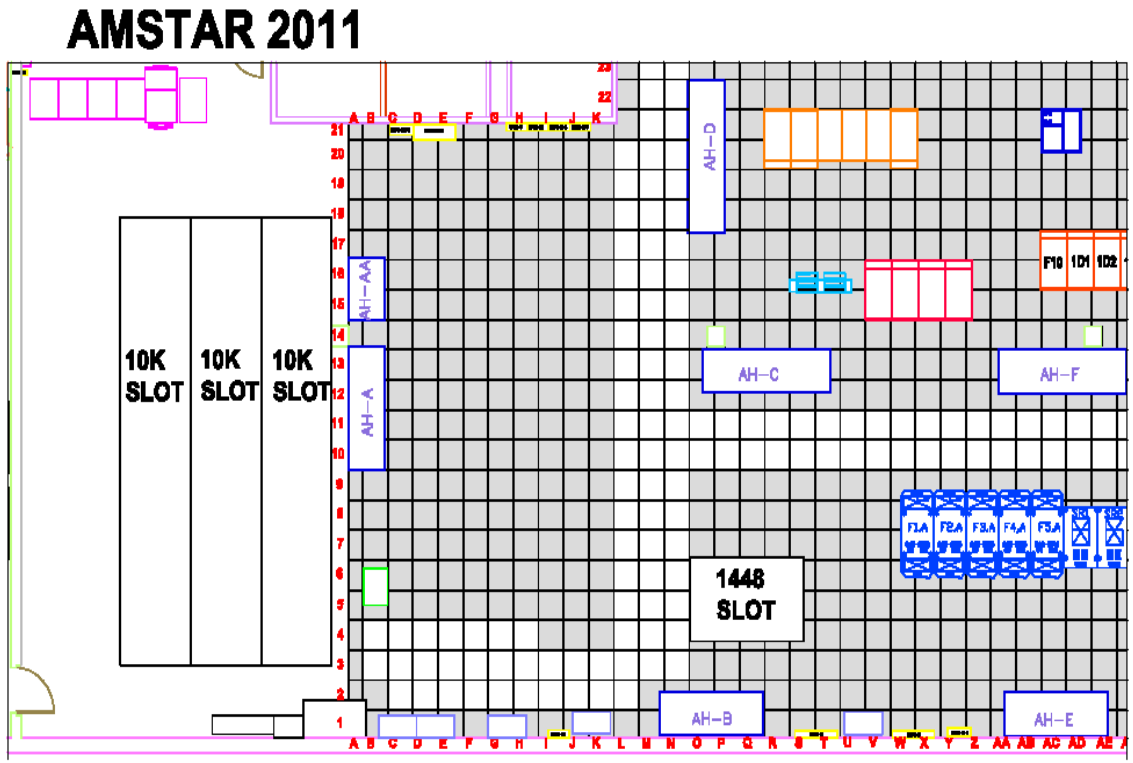
AMSTAR Progression 2010



AMSTAR Progression Sept 2011

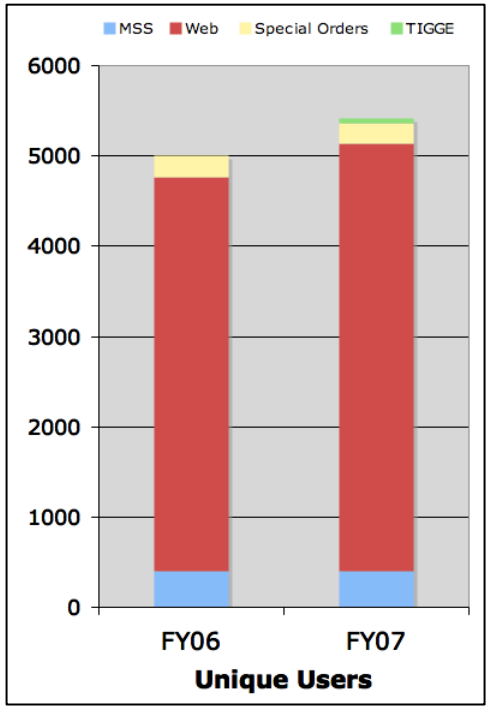
Phase 6 – 3 **Production Libraries**

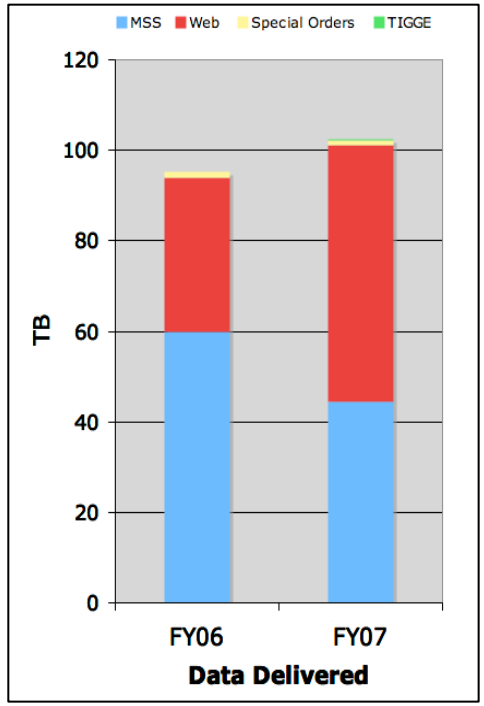
- (3) 10,000-slot SL8500 Libraries,
- (1) 1448-slot development library
- (95) T10000B tape drives
- (28,700) T10000 Tapes
- (55) T10000 cleaning tapes

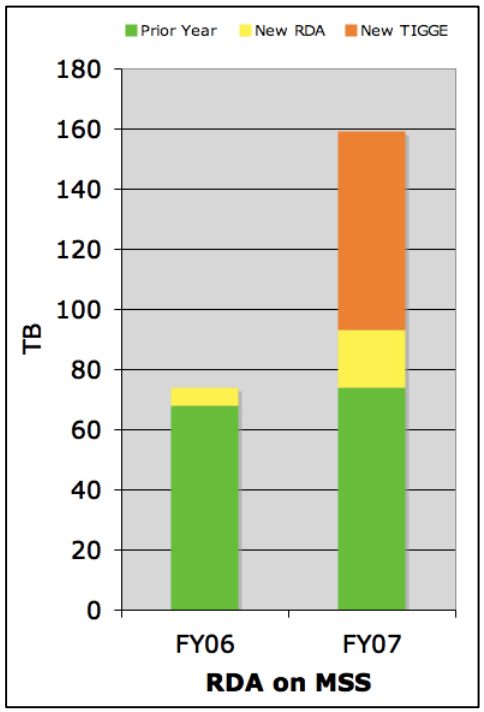


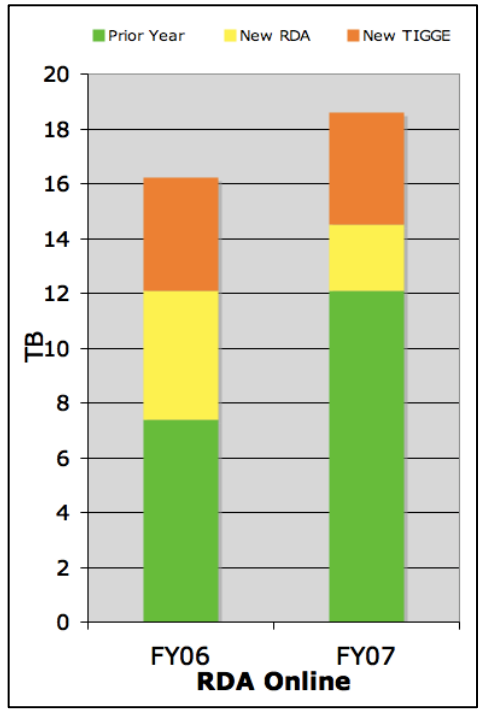
Research Data Distribution Highlights (2006/7)

- 5400 users, majority via Web (4700)
  - MSS users 400
  - Special orders 225
  - TIGGE 50
- 102 TB data delivered
- MSS growth dominated by TIGGE (66TB)
  - Other datasets increased 19 TB, up 200% from 2006
- Online availability > 18 TB









Summary of Data Providers, Oct '08

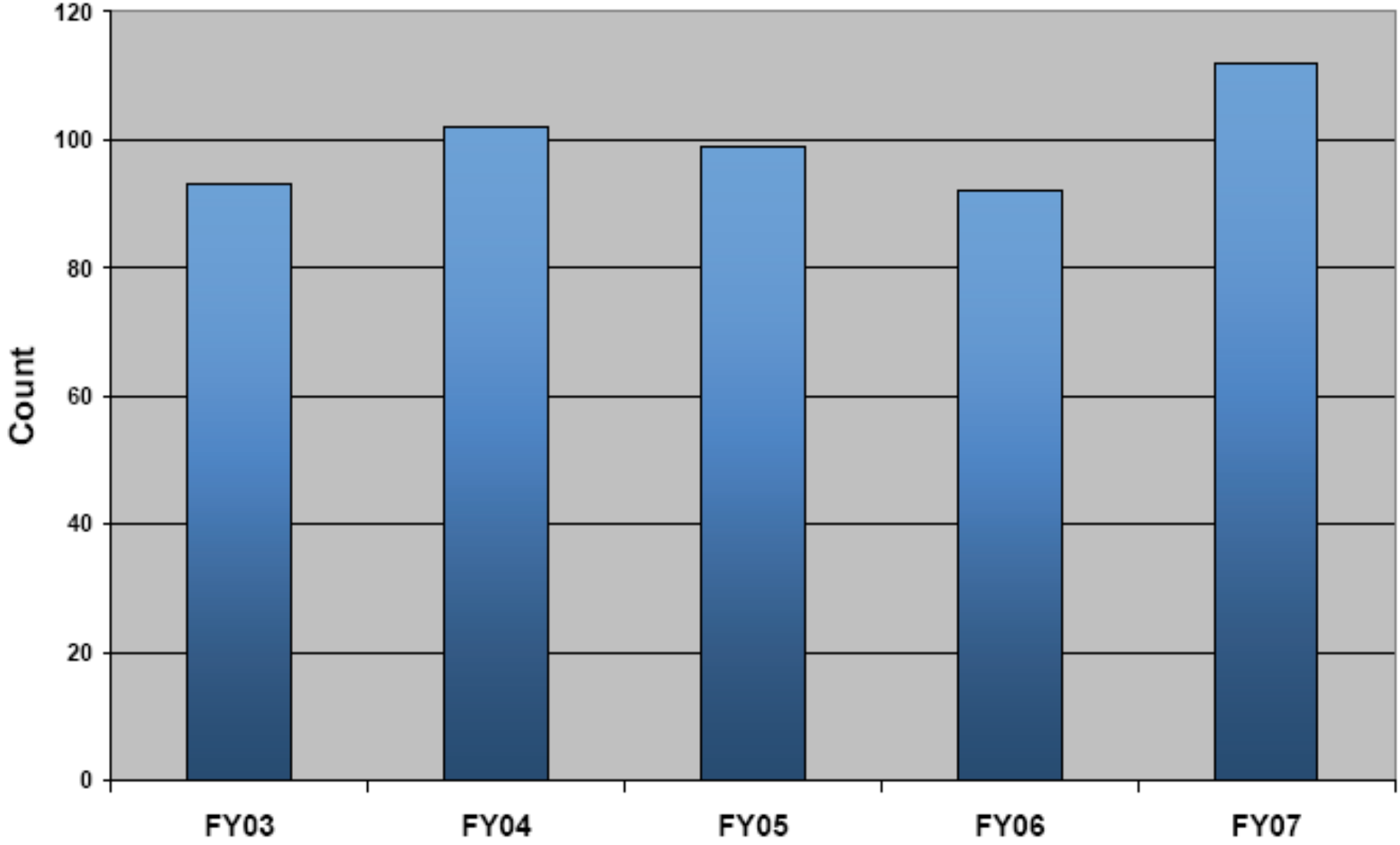
Center	Conforming Parameters	Ens. Members	Model Res.	Fcst Length	Fcsts/ Day	GB/ Day	Fields/ Day	Files/ Day
ECMWF (ecmf)	70/73	51	N200 (Reduced Gaussian)	10 day	2	115	289,734	328
ECMWF (ecmf)	70/73	51	N128 (Reduced Gaussian)	10-15 day	2	24	138,978	160
UKMO (egrr)	70/73	24	1.25 x 0.83 Deg	15 day	2	21	175,680	488
JMA (rjtd)	61/73	51	1.25 x 1.25 Deg	9 day	1	7	113,192	74
NCEP (kwbc)	69/73	21	1.00 x 1.00 Deg	16 day	4	15	371,196	1040
CMA (babj)	60/73	15	0.56 x 0.56 Deg	10 day	2	28	72,510	82
CMC (cwao)	56/73	21	1.00 x 1.00 Deg	16 day	2	8	163,674	260
BOM (ammc)	55/73	33	1.50 x 1.50 Deg	10 day	2	8	147,972	164
MF (lfpw)	62/73	11	1.50 x 1.50 Deg	2.5 day	1	.15	7,558	33
KMA (rksl)	46/73	17	1.00 x 1.00 Deg	10 day	2	5	64,124	164
CPTEC (sbsj)	55/73	15	1.00 x 1.00 Deg	15 day	2	14	97,084	244
Total					22	245	1,641,702	3,037

TIGGe Usage

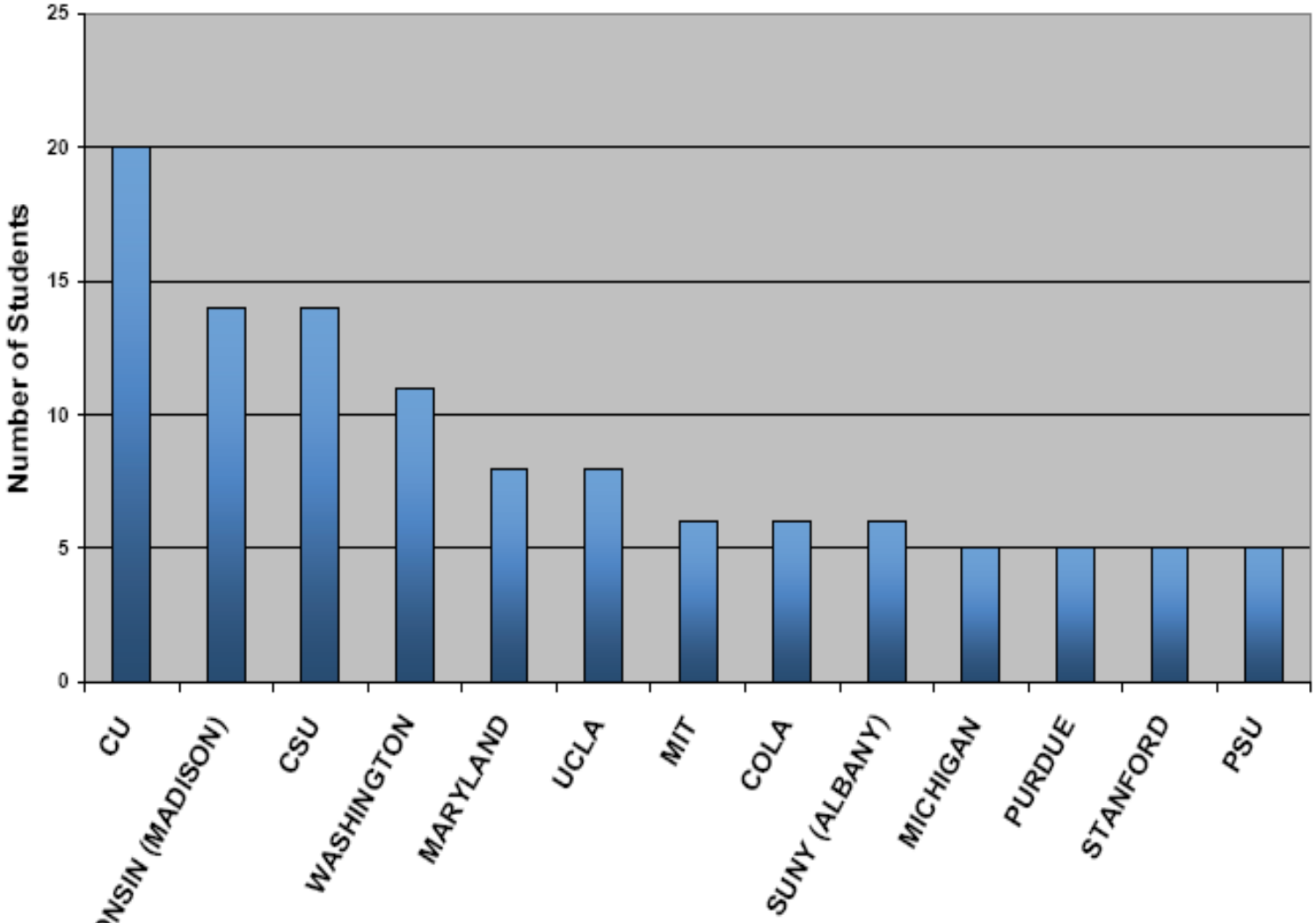
Unique Users that have downloaded data.

- **Total Number of Registered Users = 142**
- **Total volume downloaded 1.996 TB**

Universities Served by CISL



Graduate and Undergraduate Students using Computational Resources in FY07



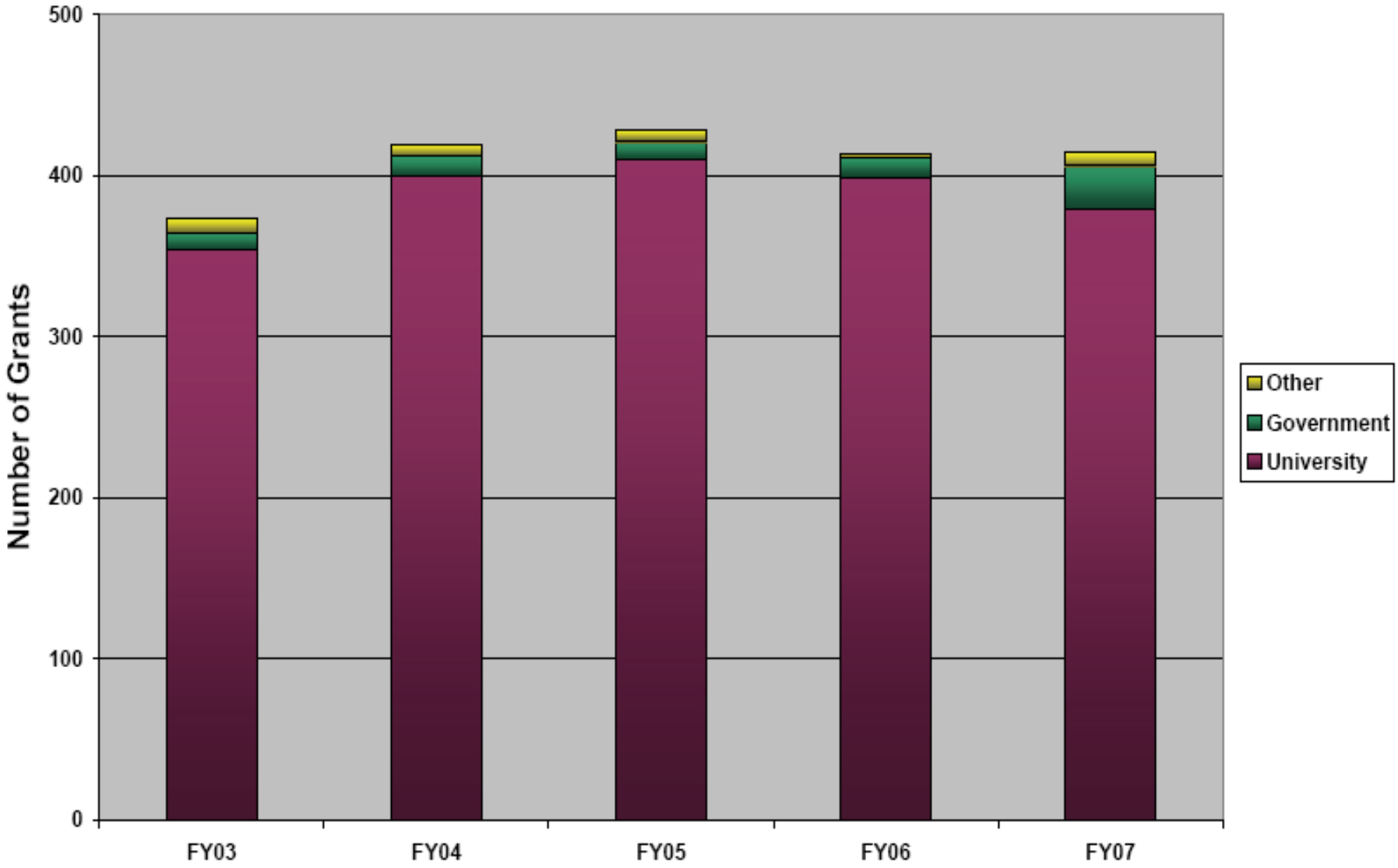
FY07

1288 Researchers Used Resouces

753 University Researchers Used Resources

291 Students Used Resources

CISL Grants to Community

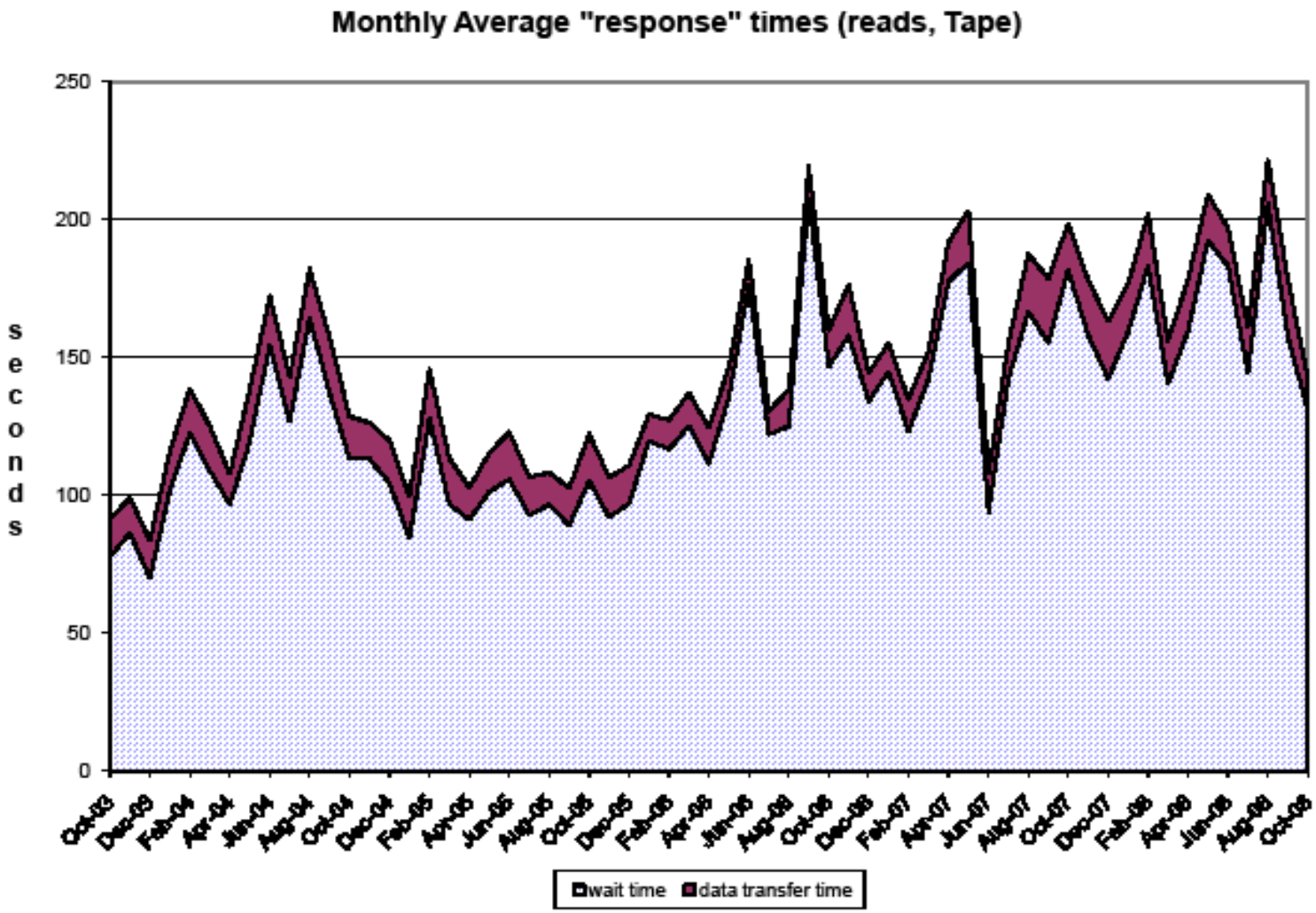


Servicing the Demand
CISL Computing Facility

- Utilization …
- … average job queue-wait times (measured in minute to hours , not days)

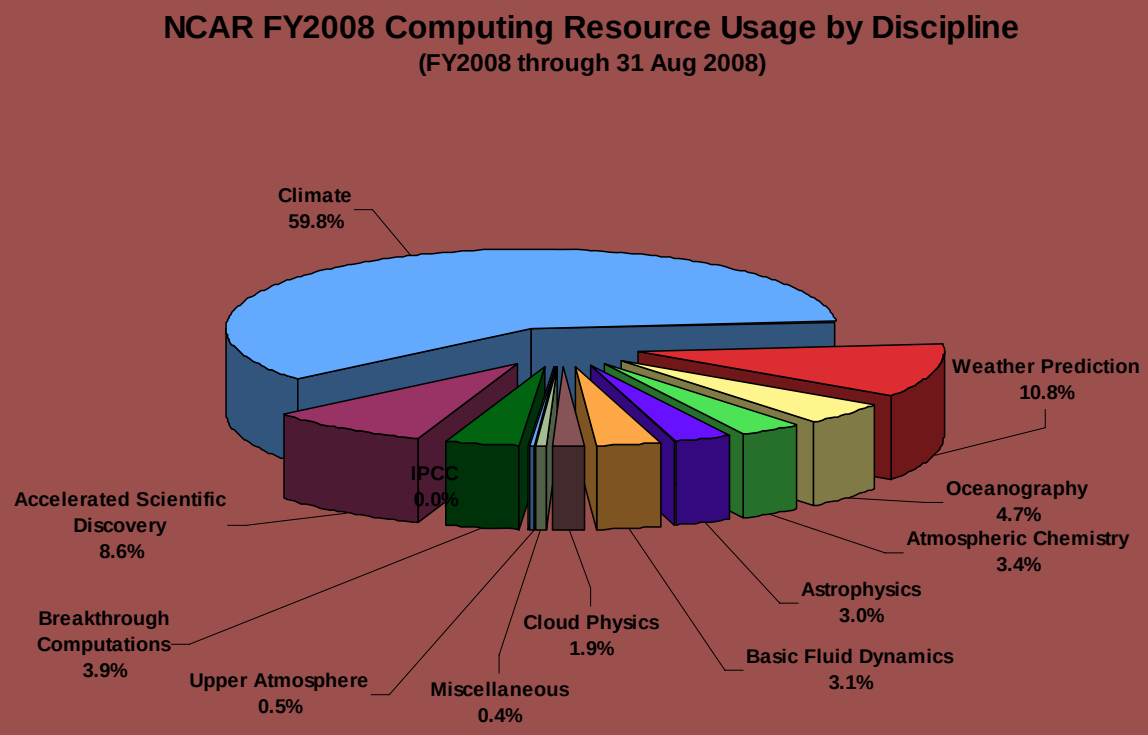
	Aug'08	2008	2007	2006	2005
Bluefire (P6)	74.9%	62.8%	-	-	-
Blueice (P5+)	-	93.5%	88.2%	-	-
Bluevista (P5)	88.1%	89.8%	89.9%	89.1%	-
Lightning(AMD)	24.6%	38.3%	47.3%	63.3%	61.5%
Bluesky 8-way LPARs (P4)	-	-	90.4%	91.7%	92.5%
Bluesky 32-way LPARs (P4)	-	-	83.3%	92.9%	94.6%

Regular Queue	Aug'08 Average Queue Wait Time	Lifetime Average Queue Wait Time
Bluefire (P6)	2m	3m
Blueice (P5+)	-	37m
Bluevista (P5)	30m	1h40m
Lightning (AMD)	0m	16m


Monthly Average "response" times (reads, Tape)
seconds
250
200
150
100
50
0
Oct-03
Dec-03
Feb-04
Apr-04
Jun-04
Aug-04
Oct-04
Dec-04
Feb-05
Apr-05
Jun-05
Aug-05
Oct-05
Dec-05
Feb-06
Apr-06
Jun-06
Aug-06
Oct-06
Dec-06
Feb-07
Apr-07
Jun-07
Aug-07
Oct-07
Dec-07
Feb-08
Apr-08
Jun-08
Aug-08
Oct-08
wait time
data transfer time

Computing Usage by Domain FY2008

- FY2008: as of 31 August '08
- Roughly 2/3 of that capacity was used for climate simulation and analysis

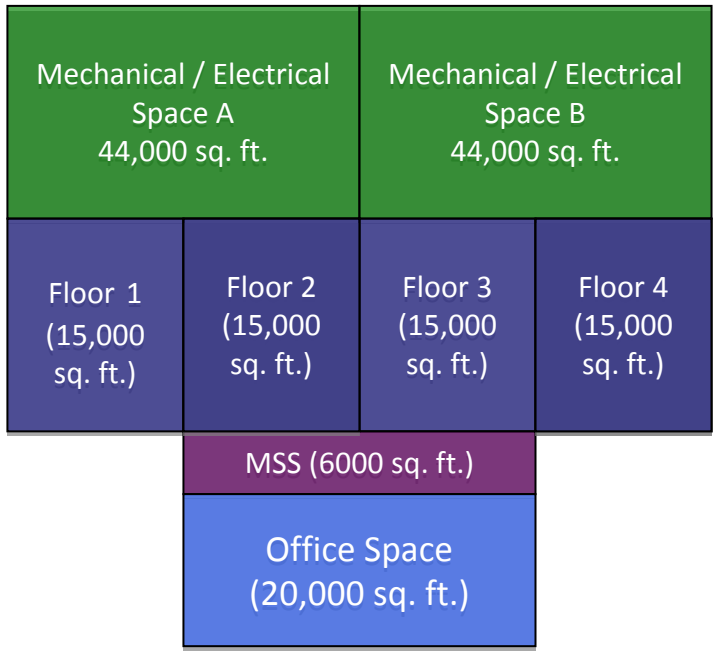


Wyoming Gov Dave Freudenthal signs Supplemental Budget Bill

March 2, 2007

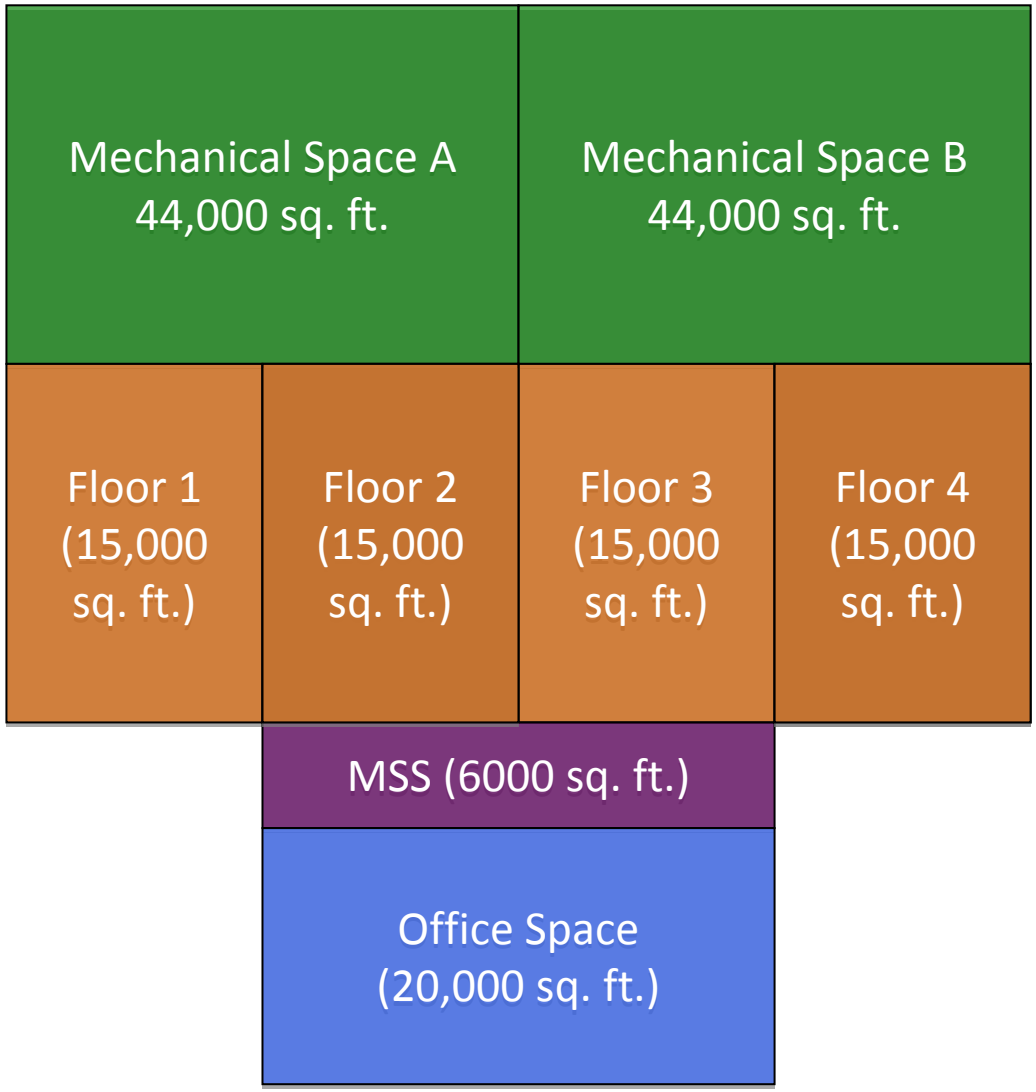
NCAR Supercomputing Center (NSC) Design

- Preferred site covers 24 acres in the North Range Business Park
- Modular facility design to be implemented, with initial size to be on the order of 100,000 sq. ft. with 15,000 sq. ft. of raised floor and 7MW
- Initial power build-out to house 4-5MW of computing



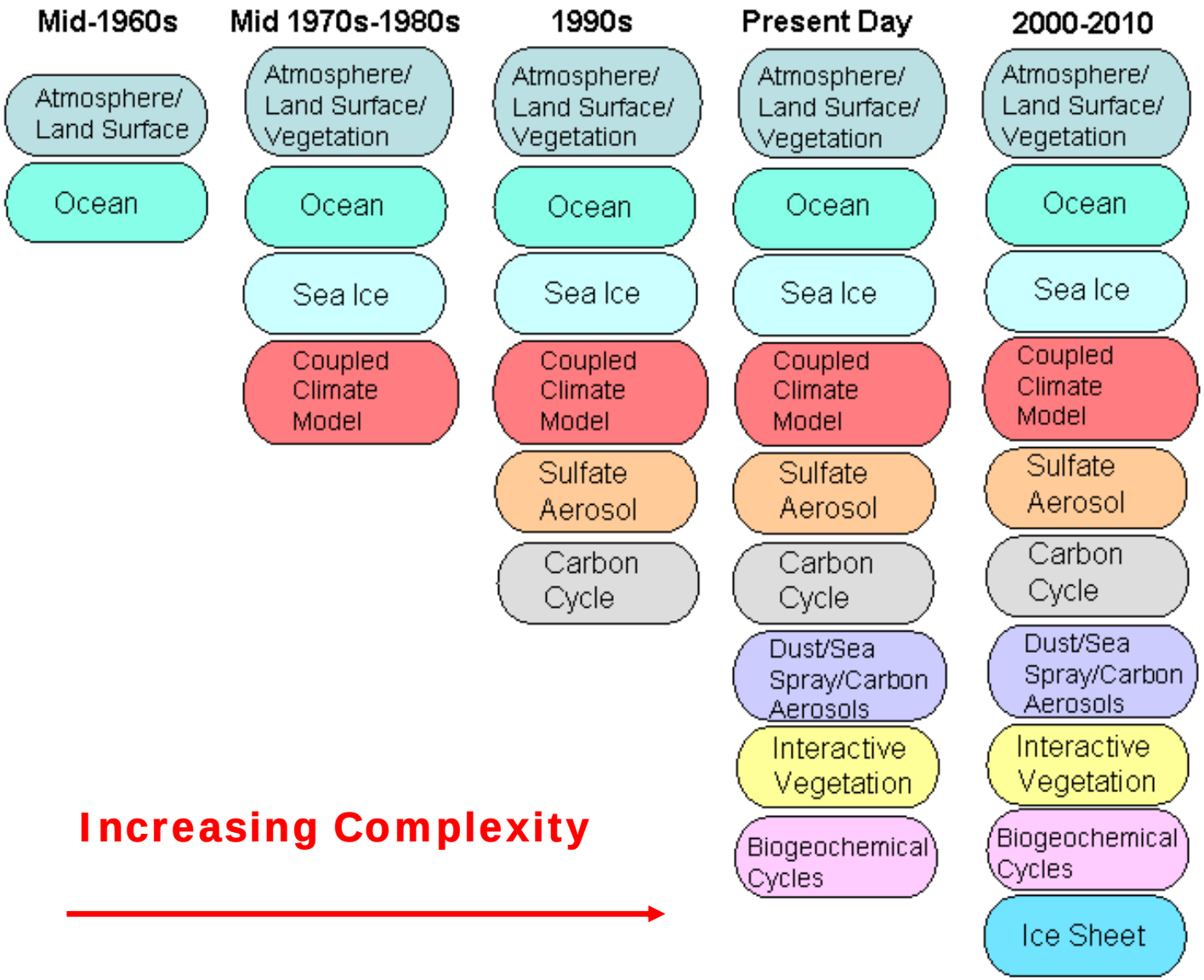
- NCAR focused on *comprehensive facility efficiency and sustainability*, including:
  - Adoption of viable energy efficient technologies to meet power and cooling needs
  - Utilization of alternative energy (wind, solar, geothermal)
  - LEED (Leadership in Energy and Environmental Design) certification

New SC Build Out

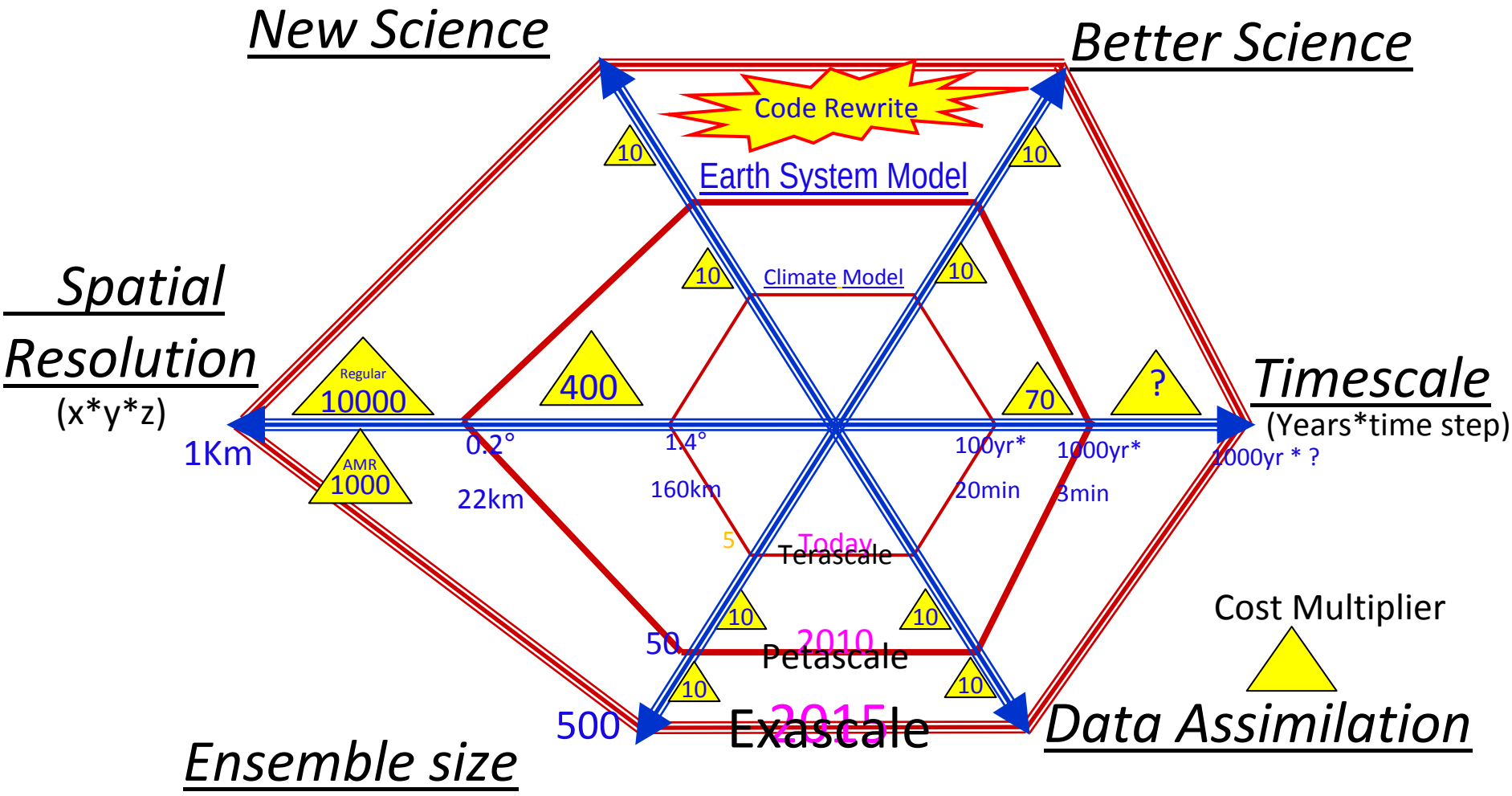


Science

Computational Requirements of Earth System Models: Complexity

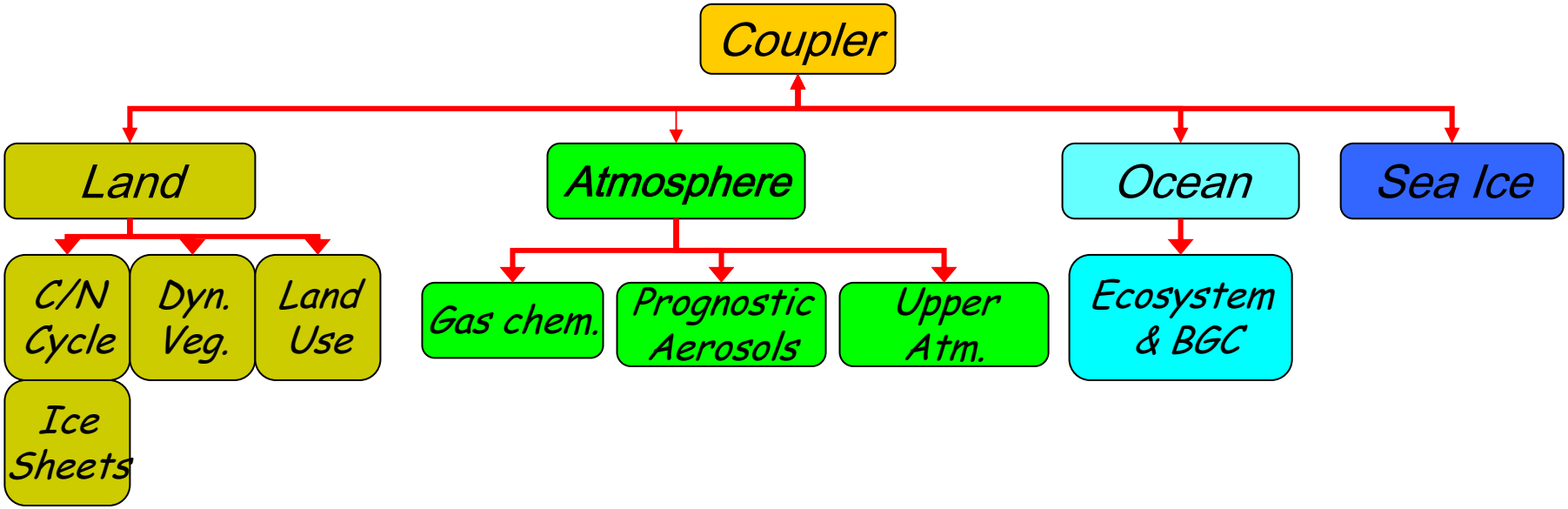


Dimensions of Climate Research



Lawrence Buja (NCAR) / Tim Palmer (ECMWF)

Climate Model Structure



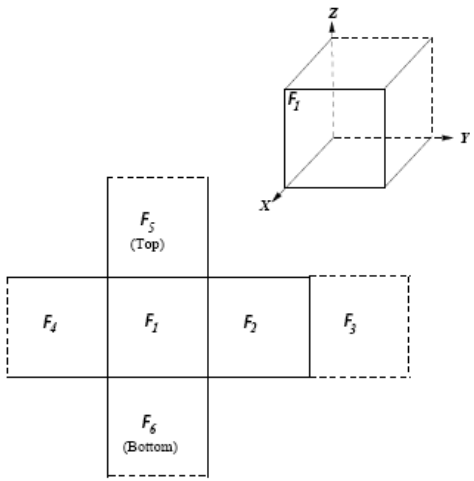
Advantages of High-Order Methods

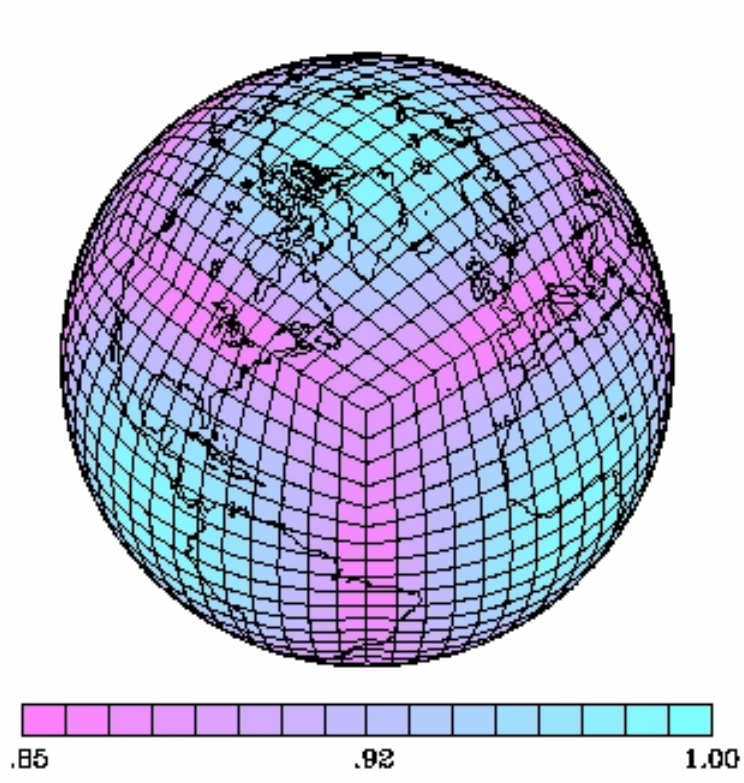
- Algorithmic Advantages of High Order Methods
  - h-p element-based method on quadrilaterals (N_e x N_e)
  - Exponential convergence in polynomial degree (N)

- Computational Advantages of High Order Methods
  - Naturally cache-blocked N x N computations
  - Nearest-neighbor communication between elements (explicit)
  - Well suited to parallel μprocessor systems

Geometry: Cube-Sphere

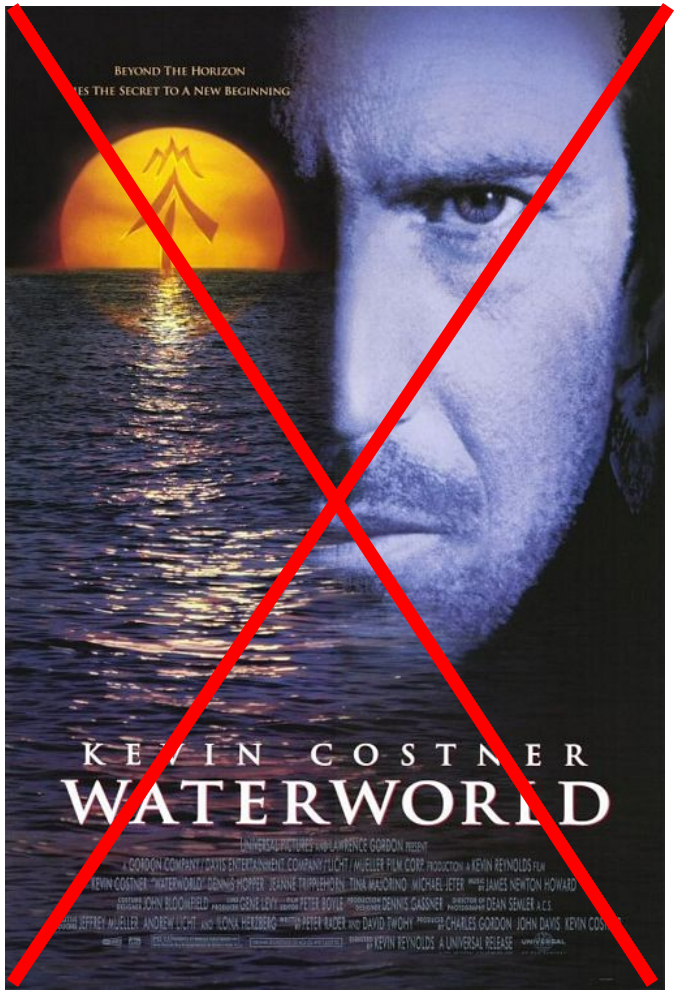
- Sphere is decomposed into 6 identical regions using a central projection (Sadourny, 1972) with equiangular grid (Rancic et al., 1996).
- Avoids pole problems, quasi-uniform.
- Non-orthogonal curvilinear coordinate system with identical metric terms





Ne=16 Cube Sphere Showing degree of non-uniformity

Validating Atmospheric Models: Aqua-Planet Experiment (APE)



- Aqua-Planet is not a bad sci-fi movie starring Kevin Costner!
- APE compares idealized climates produced by global atmospheric models on a water covered world using idealized distributions of sea surface temperature.
- APE results are used to study the distribution and variability of convection in the tropics and of mid-latitudes storm-tracks.

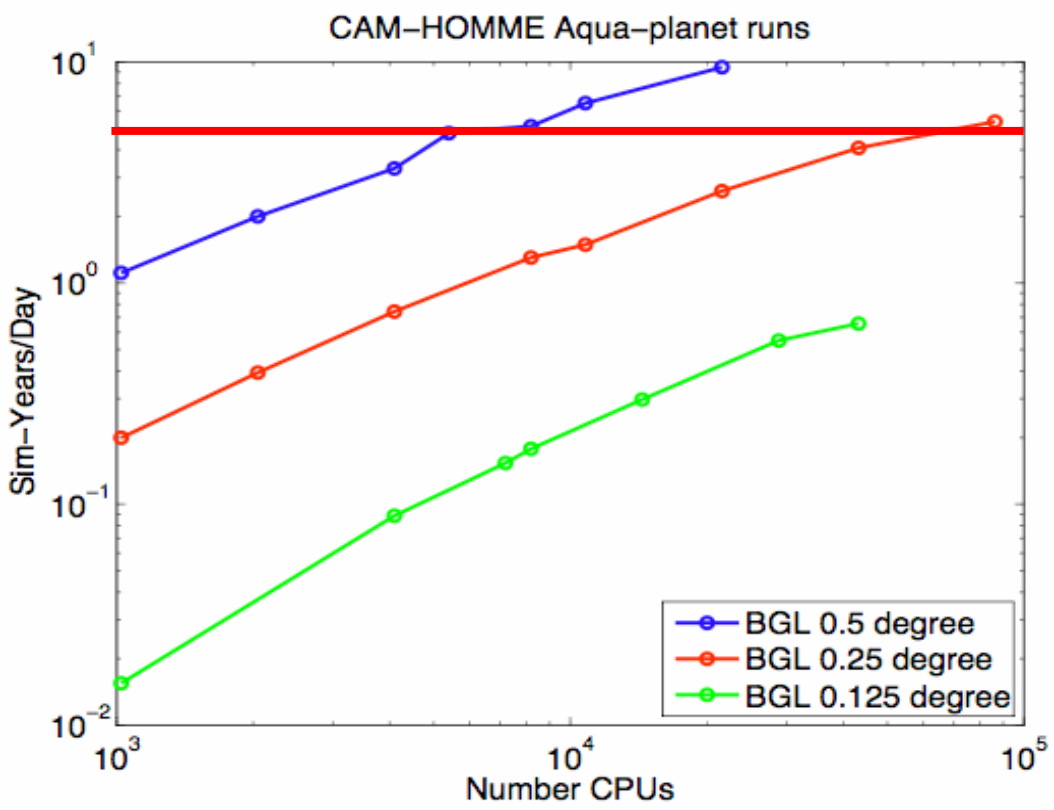
Aquaplanet: HOMME vs Eulerian CAM Performance on Globally Averaged Observables

resolution	Physics timestep (min)	Del^4 Diffusion	Precip From Convection (mm/day)	Large Scale Precip (mm/day)	Total Cloud Fraction (%)	Precipitable water (mm)
EUL T42	5	1e16	1.71	1.11	0.65	20.21
HOMME 1.9	5	1e16	1.76	1.14	0.66	20.09
EUL T85	5	1e15	1.59	1.38	0.60	19.63
HOMME 1.0	5	1e15	1.59	1.43	0.61	19.67
EUL T170	5	1.5e14	1.44	1.62	0.55	19.13
HOMME 0.5	5.5	1.5e14	1.47	1.63	0.55	19.21
EUL T340	5	1.5e13	1.36	1.75	0.50	18.75

Credit: Mark Taylor SNL and LLNL

Aqua-Planet CAM/HOMME Dycore

Full CAM Physics/HOMME Dycore
Parallel I/O library used for physics aerosol input and input data



5 years/day

Current location

Wyoming Location


Thanks
&
See you at NCAR