NCAR's Data-Intensive Supercomputing Resource:

Yellowstone

Anke Kamrath,

Director, NCAR/CISL Operations and Services

anke@ucar.edu

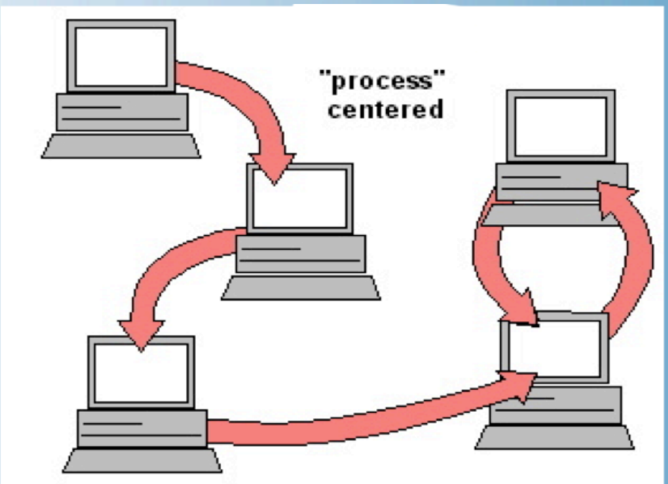
Computational & Information Systems Laboratory

CISL

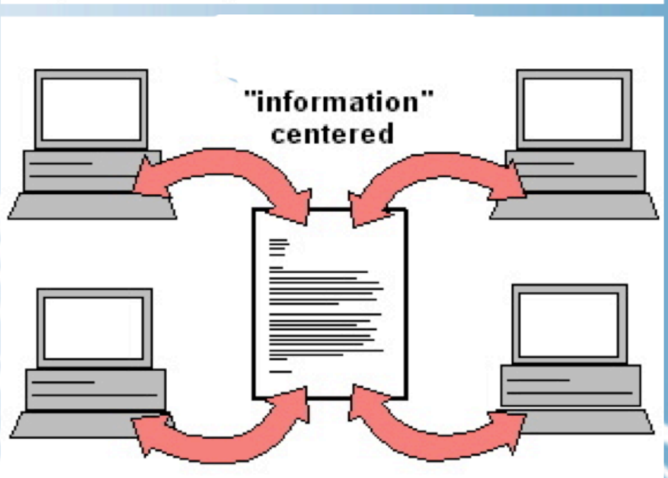
NCAR

Changing the way we do science…

"Process Centric"



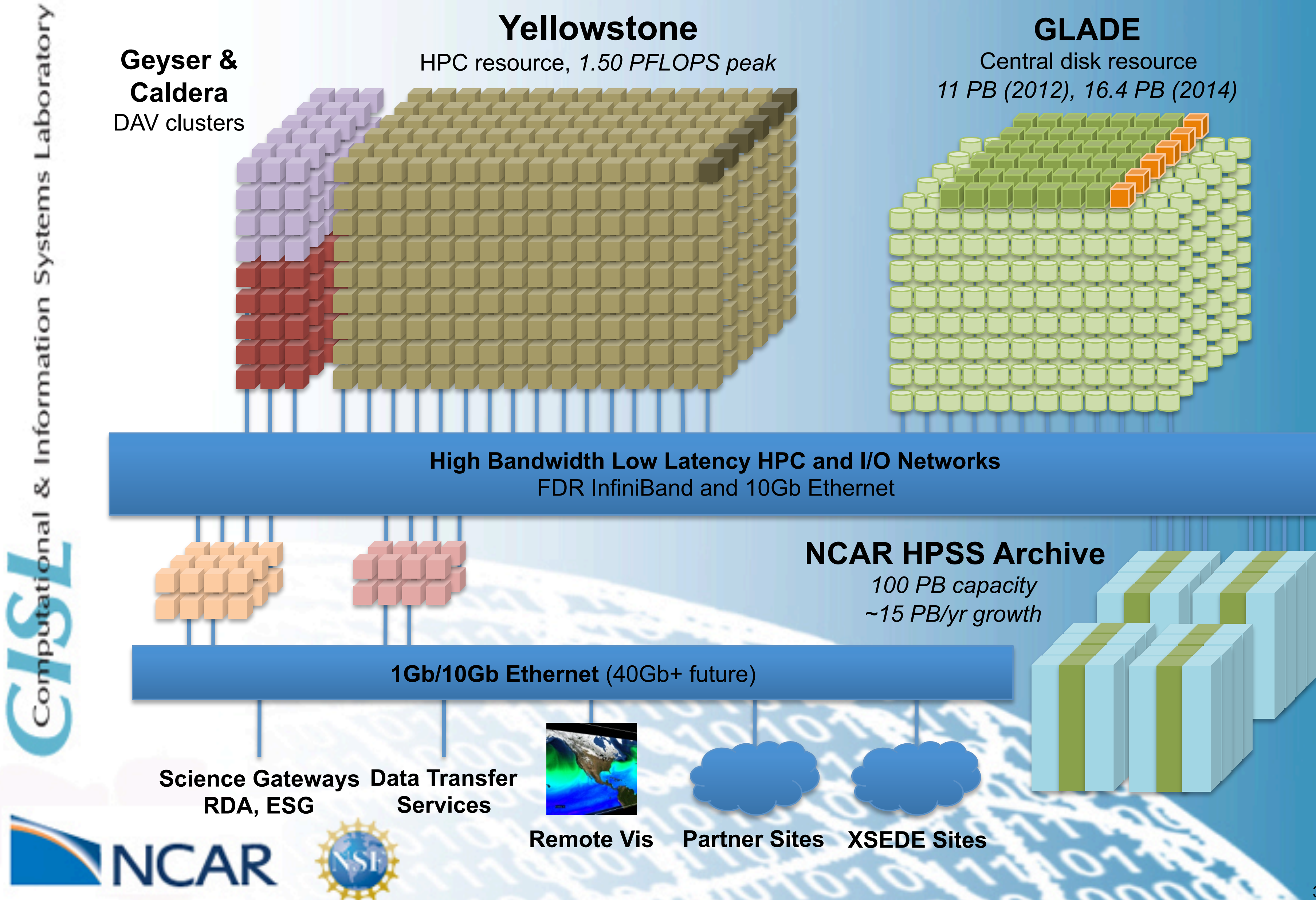
to



"Information Centric"

Computational & Information Systems Laboratory

Yellowstone Supercomputing Environment



Yellowstone

High-Performance Computing Resource

- **Batch Computation**
  - 72,288 cores total – 1.504 PFLOPs peak
  - 4,518 IBM dx360 M4 nodes – 16 cores, 32 GB memory per node
  - Intel Sandy Bridge EP processors with AVX – 2.6 GHz clock
  - 144.6 TB total DDR3-1600 memory
  - 28.9 Bluefire equivalents
- **High-Performance Interconnect**
  - Mellanox FDR InfiniBand full fat-tree
  - 13.6 GB/s bidirectional bw/node
  - <2.5 µs latency (worst case)
  - 31.7 TB/s bisection bandwidth
- **Login/Interactive**
  - 6 IBM x3650 M4 Nodes; Intel Sandy Bridge EP processors with AVX
  - 16 cores & 128 GB memory per node

CISL Computational & Information Systems Laboratory

NCAR

GLADE

(GLobally Accessible Data Environment)

- **10.94 PB usable capacity ➔ 16.42 PB usable (1Q2014)**

  Estimated initial file system sizes
  - **collections** ≈ 2 PB RDA, CMIP5 data
  - **scratch** ≈ 5 PB shared, temporary space
  - **projects** ≈ 3 PB long-term, allocated space
  - **users** ≈ 1 PB medium-term work space
- **Disk Storage Subsystem**
  - 76 IBM DCS3700 controllers & expansion drawers
    - 90 2-TB NL-SAS drives/controller
    - add 30 3-TB NL-SAS drives/controller (1Q2014)
- **GPFS NSD Servers**
  - **91.8 GB/s** aggregate I/O bandwidth; 19 IBM x3650 M4 nodes
- **I/O Aggregator Servers (GPFS, HPSS connectivity)**
  - 10-GbE & FDR interfaces; 4 IBM x3650 M4 nodes
- **High-performance I/O interconnect to HPC & DAV**
  - Mellanox FDR InfiniBand full fat-tree
  - 13.6 GB/s bidirectional bandwidth/node

Bytes/flop on current NSF HPC Portfolio

Yellowstone unique in NSF Portfolio

	TB	TF	bytes/flops
NCSA Forge	600	153	3.92
NCSA Blue Waters	25000	11500	2.17
NICS Athena	100	166	0.60
NICS Kraken	2400	1170	2.05
PSC Blacklight	150	36	4.17
TACC Lonestar4	1000	302	3.31
TACC Ranger	1730	580	2.98
SDSC Trestles	140	100	1.40
SDSC Gordon	2000	341	5.87
Total 5 centers	33120	14348	2.31
NCAR's Yellowstone Phase 1	11000	1500	**7.33**
NCAR's Yellowstone Phase 2	16400	1500	**10.93**

Yellowstone Storage – Disk & Tape

- **GLADE - Central Filesystem**
  - 11 PB (2012), growing to
  - 16.4 PB (2014)
- **HPSS – Archive**
  - 16 PB (mid-2012), growing to
  - 30-40 PB by end of 2013
- **2014 – Update Archive**
  - 100+ PB
- **Ratio Archive/Filesystem**
  - Current - 10:1
  - 2012-13 – 3:1
  - 2014 – 6:1
- **Still not enough storage…**
  - Archive and disk allocations implemented to manage demand. Need 150-200PB archive based on projections.

CISL Computational & Information Systems Laboratory

NCAR NSF

Geyser and Caldera

Data Analysis & Visualization Resource

- **Geyser: Large-memory system**
  - 16 IBM x3850 nodes – Intel Westmere-EX processors
  - 40 cores, **1 TB memory,** 1 NVIDIA GPU *per node*
  - Mellanox FDR full fat-tree interconnect
- **Caldera: GPU computation/visualization system**
  - 16 IBM x360 M4 nodes – Intel Sandy Bridge EP/AVX
  - 16 cores, 64 GB memory per node
  - 2 NVIDIA GPUs per node
  - Mellanox FDR full fat-tree interconnect
- **Knights Corner system (Q2 2013 delivery)**
  - Intel Many Integrated Core (MIC) architecture
  - 16 IBM Knights Corner nodes
  - 16 Sandy Bridge EP/AVX cores, 64 GB memory
  - 1 Knights Corner adapter per node
  - Mellanox FDR full fat-tree interconnect

CISL Computational & Information Systems Laboratory

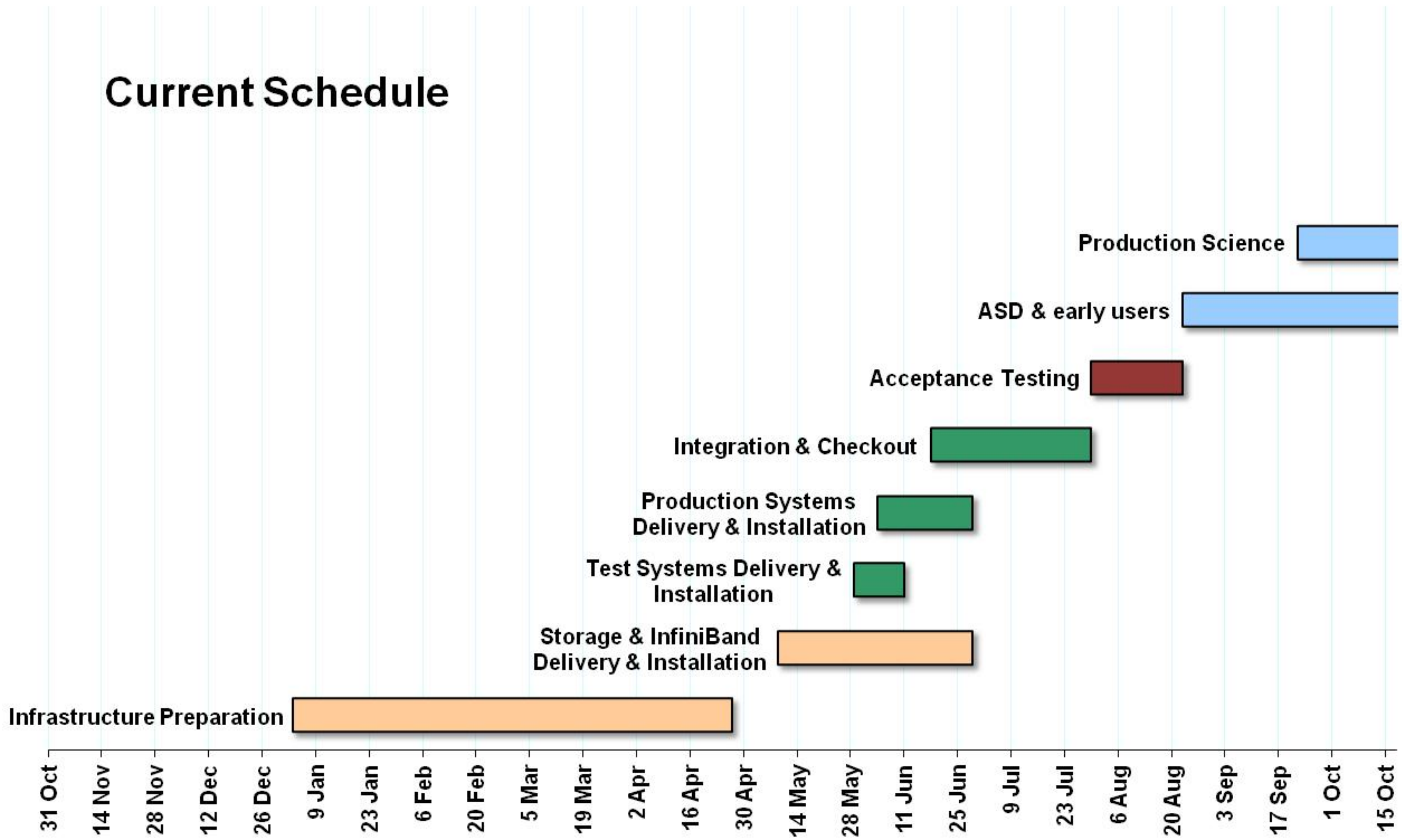
NCAR NSF

Yellowstone Power Efficiency vs Bluefire

	Yellowstone	Bluefire
Processor	2.6 GHz Xeon E5	4.7 GHz POWER6
Total Batch Processor Cores	72,288	3,744
Batch portion peak TFLOPs	1500	72
Power Consumption	~1.9 MW	540 kW
Watts/peak GFLOP	1.3	7.5
Peak MFLOP/Watt	800	133
Average workload floating point efficiency	5.4% (estimate)	3.9% (measured)
Sustained MFLOPs/Watt (on NCAR workload)	~43	~6
Bluefire-equivalents	28.9	1

For 3.5x more power, Yellowstone delivers 28.9x more computational performance than Bluefire.

Yellowstone Schedule



CISL Computational & Information Systems Laboratory

NCAR

NWSC Installation

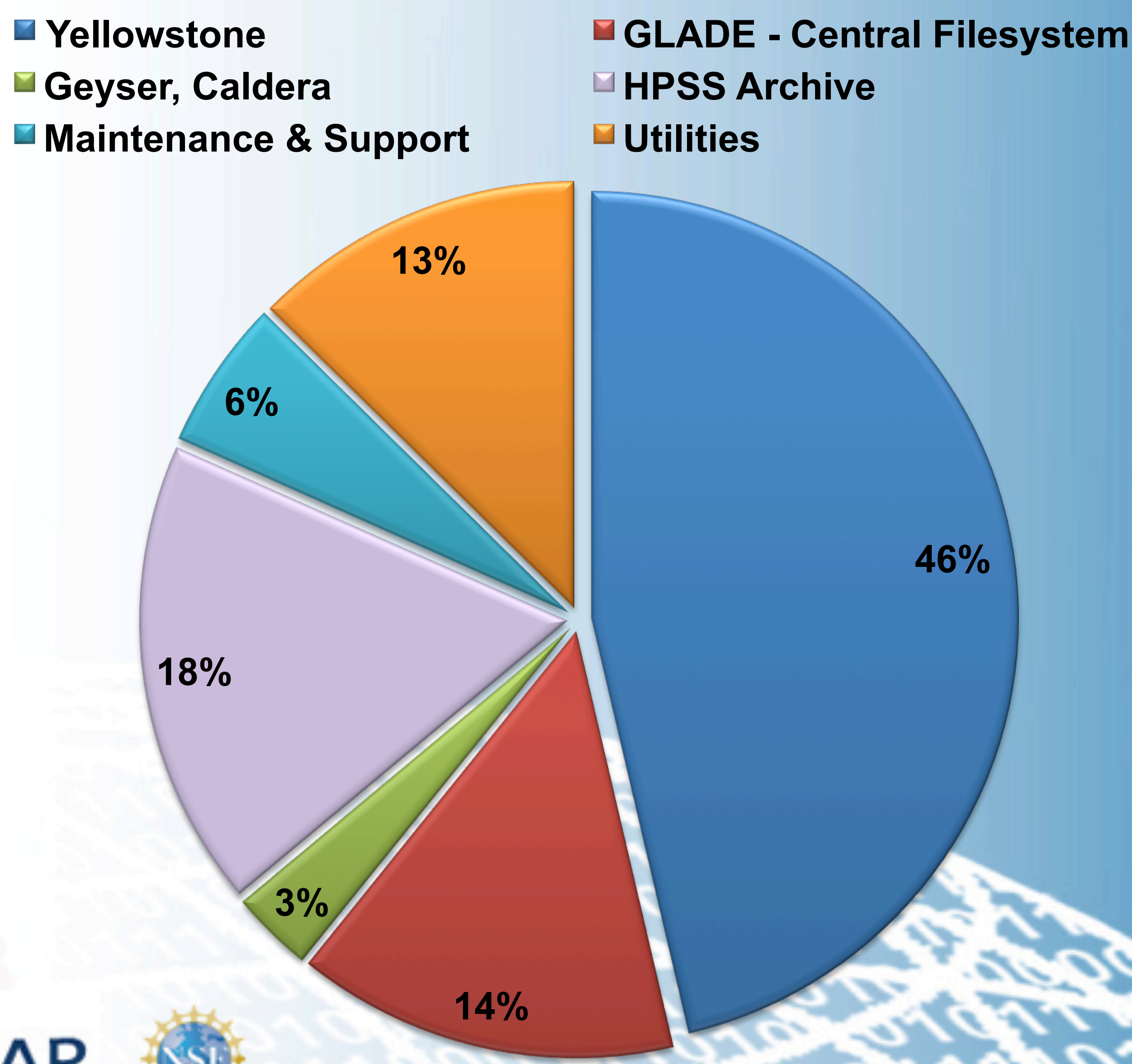
CISL Computational & Information Systems Laboratory

NCAR NSF

When can users get going on this?

- **Yellowstone in Acceptance in August:**
  - If all goes well, ready for users in **September**
- **Until its ready – we have:**
  - Bluefire (~4-8 week overlap once Yellowstone in production)
  - Janus Cluster (similar to Yellowstone)
    - 16,416 cores total – 184 TFLOPs peak
    - 1,368 nodes – 12 cores, 24 GB memory per node
    - Intel Westmere processors – 2.8 GHz clock
    - 32.8 TB total memory
    - QDR InfiniBand interconnect
    - Red Hat Linux
    - Deployed by CU-Boulder in collaboration with NCAR
    - ~10% of the system allocated by NCAR
    - *Small allocations to university, NCAR users*
    - CESM, WRF ported and running
    - Key elements of NCAR software stack installed
    - www2.cisl.ucar.edu/docs/janus-cluster

Yellowstone Environment Lifetime Costs



Many challenges ahead …
(post-Yellowstone)

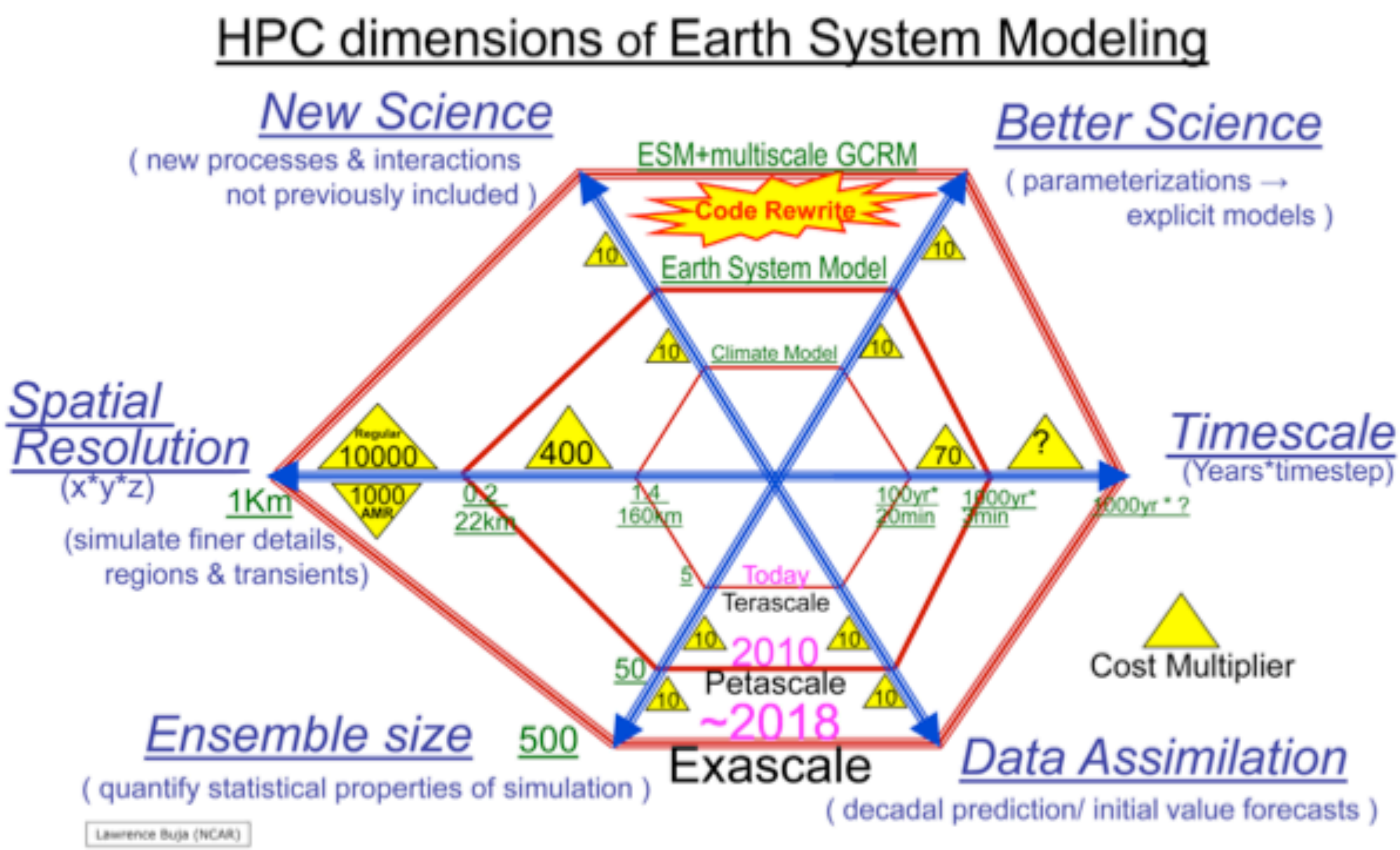
- **Follow-on System (~2015) Likely to Have:**
  1. Xeon and Many-Cores (e.g., nVIDIA, MIC)
     - Need for improved science-FLOPS/watt.
  2. Less Memory/Core
     - Number of cores outpacing improvement in memory pricing
     - Need to live in smaller per core memory footprint (i.e., ~1 GB/core as compared with 2 GB/core on yellowstone)
  3. Budget constrained storage environment
     - Cost of storage capacity/bandwidth is outpacing compute
- **We need to prepare now for these Challenges**

Historical FLOP/Power Efficiency on NCAR Systems

Name	Model	Peak GFLOPs	Sus GFLOPs	Power (kW)	Sus MFLOP/ Watt	Watt/ Sus GFLOP	Est'd Power Cost/yr
chipeta	CRI Cray J90se/24	4.8	1.58	7.5	0.21	4753	$5,625
ute	SGI Origin2000/128	64	7.85	51.1	0.15	6513	$38,325
blackforest	IBM SP/1308 (318) WH2/NH2	1,962	121.6	140.0	0.87	1151	$105,000
bluesky	IBM p690/32 (50) Regatta-H/Colony	8,320	343.6	415.0	0.83	1208	$311,250
lightning	IBM e325/2 (128) Opteron Linux cluster	1,144	63.8	48.0	1.33	753	$36,000
bluevista	IBM p575/8 (78) POWER5/HPS	4,742	347.6	210.6	1.65	606	$157,950
blueice	IBM p575/16 (112) POWER5+/HPS	13,312	1,000.2	325.4	3.07	325	$244,050
Bluefire (2008)	**IBM Power 575/32 (128) POWER6 DDR-IB**	**77,005**	**2,987.8**	**538.2**	**5.55**	**180**	**$403,654**
Frost (2009)	IBM BlueGene/L (4096/2)	22,938	741.5	83.1	8.92	112	$62,325
lynx	Cray XT5m (912/76)	8,130	487.8	35.0	13.9	72	$26,250
Yellowstone (2012)	**IBM iDataPlex/FDR-IB**	**1,503,590**	**80,950**	**1,900**	**42.6**	**23**	**$1,700,000**
next system??	*Possible - Intel Xeon only estimates ->*	*10,000,000*	*500,000*	*5,000*	*100.0*	*10*	*$3,750,000*
next system??	*Possible - Intel Xeon & GPUs estimates ->*	*40,000,000*	*500,000*	*3,000*	*166.7*	*6*	*$2,250,000*

What does a new supercomputer enable?

Expanding the modeling envelope



Inefficient Use of Memory

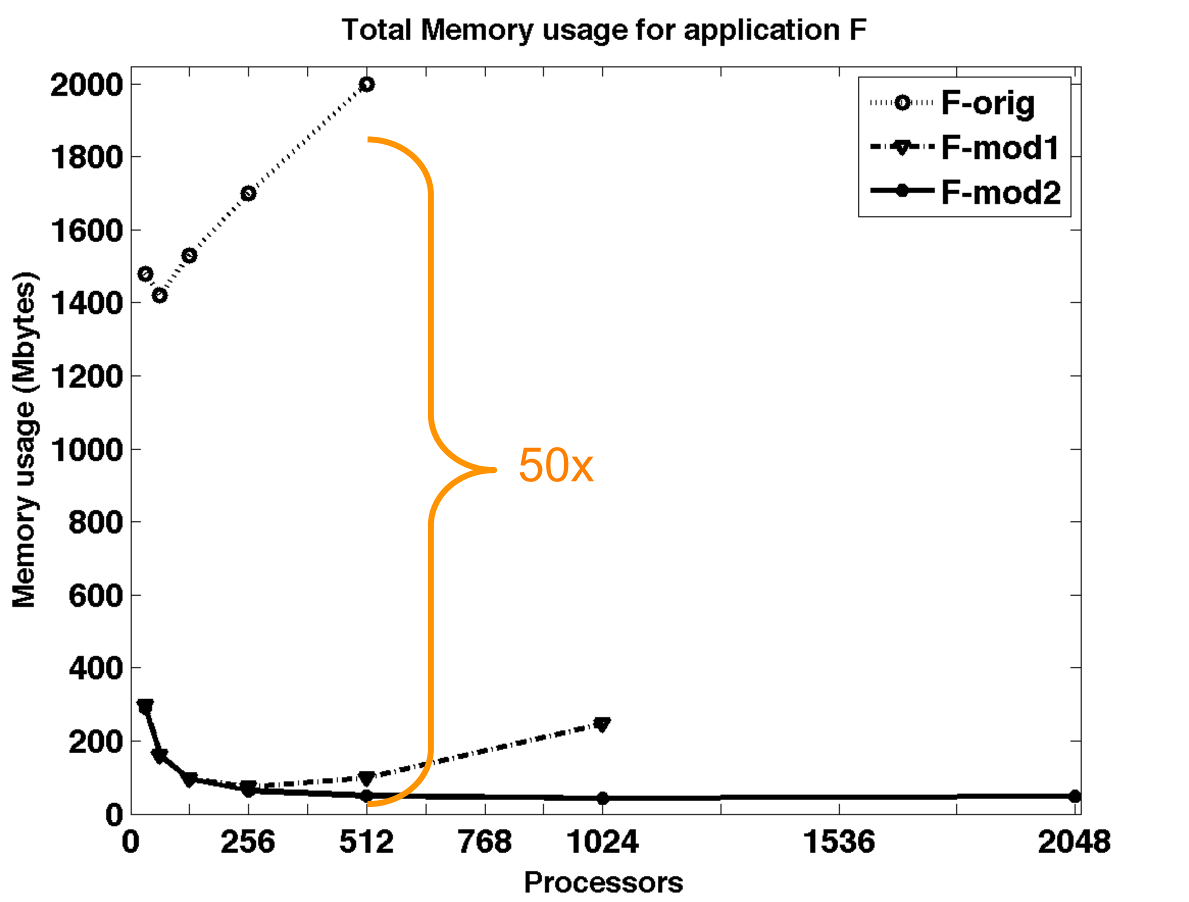
- **Replicated Metadata**
  - Describes the location of something else
    - eg: message passing schedule, domain decomposition
  - Consider: p_i sends 'n' bytes to p_j
    - Don't store information about p_j, p_i on p_k if k != i,j
  - High resolution version of CLM on 10K processors --> 29 TB
  - Ignorance is bliss! :-)
- **Excessive Global Arrays**
  - global arrays: an array the size of the entire computational mesh
  - Persistent versus temporary global arrays
  - Low res ---> no big deal; High res --> can be fatal !
  - Example: CLM
    - Original: ~500
    - Now: 1 (temporary)

CISL Computational & Information Systems Laboratory

NCAR NSF

Using less memory to get same science done

Total memory usage for CLM



Improving Data Workflow and Management

- **Today:**
  - We are seeing many challenges with Data Analysis of CMIP5 data
    - Atomic NCO commands very inefficient to transform the raw CESM output into time series. Prevents use of buffering
    - NCAR GLADE GPFS 2MB Blocks, NETCDF3 10-150kB used – very inefficient. 10-20X too much data being moved.
    - Overall workflow complex and inefficient, and lacking automation
- **Work ahead:**
  - Need improvements on all fronts
    - More efficient data management, scripts,
    - Thinking carefully about what you store, what should be done during the simulation phase rather than after the fact
    - Tuning of systems to better support workload
    - Much more…
  - Nearly unworkable "today" – won't survive at all in future.

In Conclusion

- Excited about new science that will be enabled with Yellowstone!!
- However, a lot of work ahead to prepare for Yellowstone follow-on….

CISL Computational & Information Systems Laboratory

NCAR NSF

Questions

Yellowstone Software

- **Compilers, Libraries, Debugger & Performance Tools**
  - **Intel** Cluster Studio (Fortran, C++, performance & MPI libraries, trace collector & analyzer) 50 concurrent users
  - **Intel** VTune Amplifier XE performance optimizer 2 concurrent users
  - **PGI** CDK (Fortran, C, C++, pgdbg debugger, pgprof) 50 conc. users
  - **PGI** CDK GPU Version (Fortran, C, C++, pgdbg debugger, pgprof) for DAV systems only, 2 concurrent users
  - **PathScale** EckoPath (Fortran C, C++, PathDB debugger) 20 concurrent users
  - Rogue Wave **TotalView** debugger 8,192 floating tokens
  - **IBM** Parallel Environment (POE), including IBM HPC Toolkit
- **System Software**
  - **LSF-HPC** Batch Subsystem / Resource Manager
    - IBM has purchased Platform Computing, Inc., developers of LSF-HPC
  - Red Hat Enterprise **Linux** (RHEL) Version 6
  - IBM General Parallel Filesystem (**GPFS**)
  - Mellanox Universal Fabric Manager
  - IBM xCAT cluster administration toolkit