Research and Development Projects in CISL

Rich Loft
Director,
Technology Development Division, CISL
October 6, 2011

NCAR

Janus supercomputer
NWSC and the exascale

NCAR

Janus Supercomputer

- Joint project between CU Boulder, NCAR and CU Denver
- Number 52 on the June 2011 top 500 list with 152.2 teraflops on Linpack
- Funded in part through a NSF MRI project, PI/CO-I team includes
  - Henry Tufo – CU, Boulder/NCAR
  - Jan Mandel – CU, Boulder, Denver
  - James Syvitski - CU
  - Richard Loft -NCAR
  - Keith Julien – CU, Boulder

NCAR

Hardware -Janus Supercomputer

- 1368 compute nodes (Dell C6100)
  - two, 2.8 GHz, 6 core Intel Westmere processors per node
  - 2 GB/core; 24 GB per node
  - 16,428 total cores
  - One QDR NIC per node
- Fully non-blocking QDR Infiniband network
- 960 TB of usable Lustre-based scratch storage
  - 16-20 GB/s max throughput
- No local storage on the compute nodes
- No battery backup of the compute nodes

NCAR

CU Boulder facility

- "Containerized" solution
  - 65'x35' fabricated to order data center
  - 15 year lifespan
- Power
  - 2 MW feed
  - 60 kw N+1 UPS for storage, login and core network
- Cooling
  - Evaporative cooler, "free-cooling" flat plate
  - 337 ton chiller (projected to run 5% of overall time
- Facility target PUE: 1.2 ±10%

NCAR

Available storage

- Home directories
  - /home/thha0714
  - 2 GB Quota
- Parallel scratch on Janus
  - /lustre/janus_scratch/thha0714/
  - No quota but usage is monitored

NCAR

Environment

- Email rc-help@colorado.edu to report any problems
- Log into login.rc.colorado.edu
- RHEL 5.3
- Load software with Dotkit
  - use –l (lists all available packages)
  - use ICS (intel 12.0 compilers)
  - use OpenMPI-1.4-ICS
- Build your own software/tools
- Submit jobs with qsub (torque/maui/moab)
- Documentation: https://www.rc.colorado.edu/crcdocs

NCAR

Software Dotkit configurations Supported on Janus

OpenMPI 1.4.3	MPICH2-1.4	NetCDF 4.1.3	Parallel NetCDF 1.2	HDF 4.2.6	HDF5 1.8.7
X			X		
		X		X	X
X		X		X	X
	X	X		X	X
X		X	X		

NCAR

Queues

- janus-debug: 1 hour wall time, up to 512 cores
- janus-short: 4 hour wall time, up to 1024 cores
- janus-normal: 24 hour wall time, up to 1024 cores
- janus-wide: 24 hour wall time, up to 5120 cores
- janus-long: 168 hour wall time, up to 960 cores

NCAR

NCAR's Share of Janus

- 9.8% of the resource
  - 14.1 M Janus core hours (JCH) per year
  - 94 TB of disk
- Conversion: 1.4 GAU/JCH
- Portion made available for BF overflow:
  - 1.7 MJCHs -> 2.38 MGAUs
- Data transfer via dedicated 10 Gb/sec link
- PI's must have GAU allocation for storage

NCAR Janus Allocation Requests

	allocated core hours	requested disk (TB)	allocated disk (TB)
SM. DISC	180,000	1.5	1.5
LG. DISC	2,000,000	30	15
SM. MRI	100,000	0.5	0.5
LG. MRI	5,740,000	54.8	18.8
LG. COMMUNITY	1,174,000	85	85
SM. COMMUNITY	100,000	0.5	0.5
TOTAL	9,294,000	172.3	38.3

Currently remaining for May, 2011 allocation: 4.75M

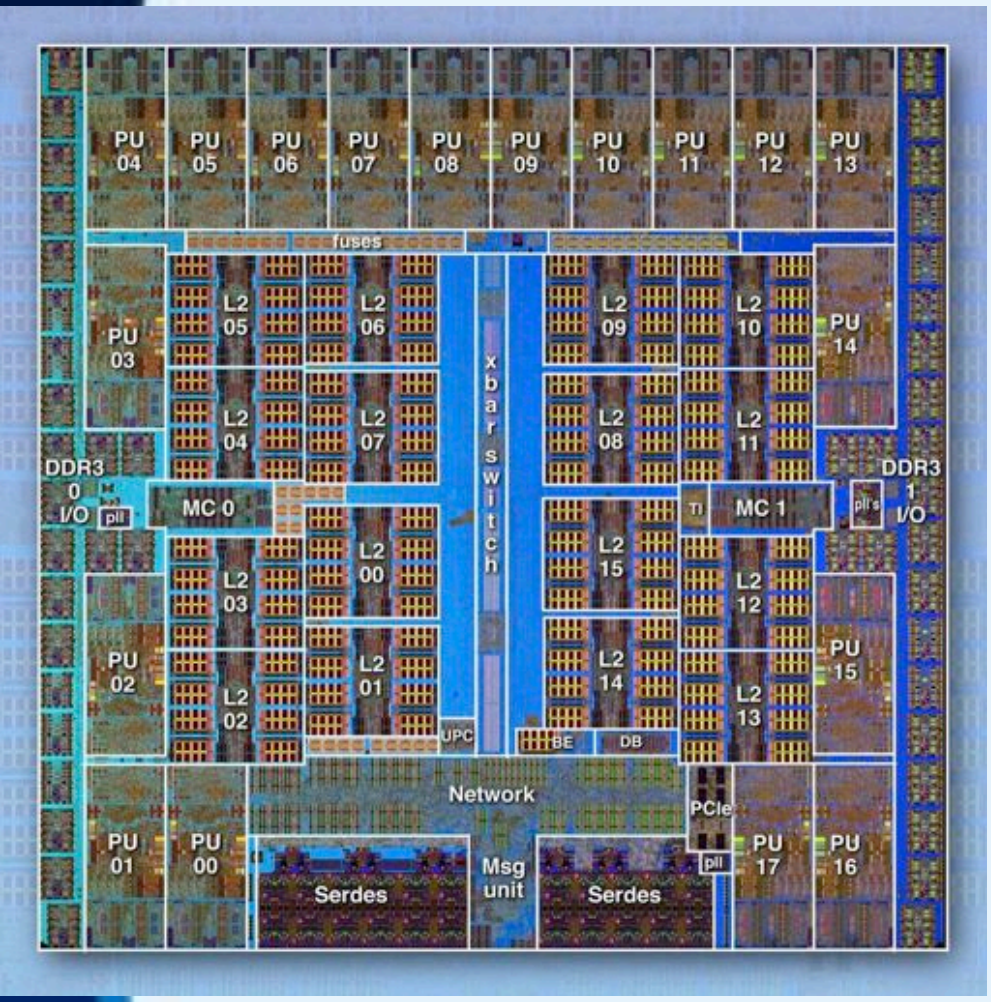
NCAR

How close to the Exascale can NWSC get by ~2018?

Assume both A & B rooms "full": 8 MW system

NCAR

Blue Gene/Q: System on a Chip



Blue Gene/Q chip layout

Vital Statistics

- 45 nm technology
- clocked @ 1.6 GHz
- Quad FPU's (8 flops/clock tick)
- 16 user cores + 1 system core + 1 spare
- 128 flops/clock tick
- 204.8 GFLOPS/chip
- 55W/chip
- 268 pJ/flop
- 42.7 GB/sec

Extrapolate to exascale-

- 1 exaflops -> 4.9 M chips
- 268 MW from chips
- 108 MW from memory
- in 4800 racks
- program ~78 million cores
- 208 petabytes/sec peak memory BW.

And this could be considered the Prius of modern day multi-processor systems!

NCAR

Semiconductor scaling was a beautiful thing

- Scaling should allow us to keep shrinking transistors from 44, 22 and 11 nm with the same power footprint…
- Putting 4x and 16x transistors on the same silicon area ($mm^{2)}$
- Requires power per transistor to scale as well
  - By 1/4 and 1/16…
- But there is a problem: this is getting harder
- we're approaching the atomic scale.
- The end of scaling.

The Dark Silicon Problem

“ . . . a 11nm process technology could deliver devices with 16 times more transistors . . . but those devices will use a third as much energy as today’s parts, leaving engineers with a power budget so pinched they may be able to activate only nine percent of those transistors.”

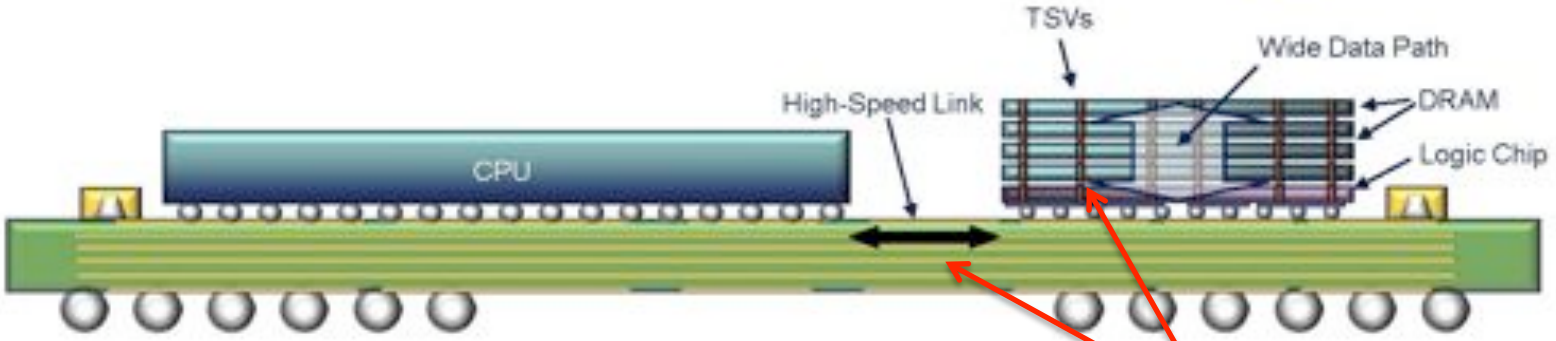
1) Better make sure the “right” transistors are active!
2) So ~90 MW/exaflops if this view holds.

NCAR

Micron's 3D-Stacked Memory Prototype Solution

HMC Near Memory – MCM Configuration

- All links are between host CPU and HMC logic layer
- Maximum bandwidth per GB capacity

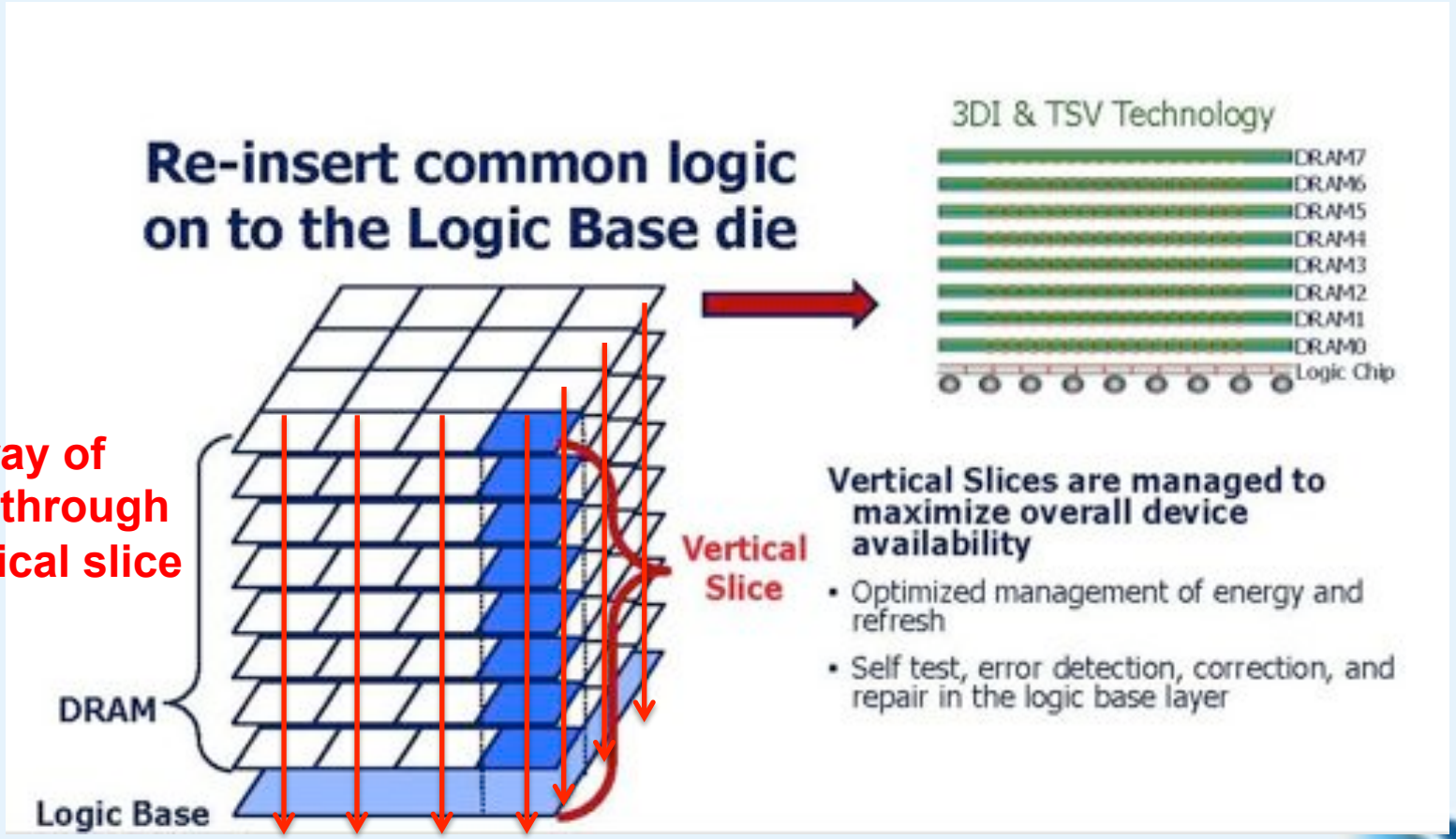


Notes: MCM = multi-chip module
Illustrative purposes only; height is exaggerated

Shorter wires = less power!

Reference: J T Pawlowski of Micron @ Hot Chips 23 August 2011

Transposing Memory in HMC: Higher memory bandwidth, lower latency



Reference: J T Pawlowski of Micron at Hot Chips 23, August 2011

NCAR

HMC: Memory Power Efficiency

Technology	VDD	IDD	BW GB/s	Power (W)	mW/GB/s	pj/bit	real pJ/bit
SDRAM PC133 1GB Module	3.3	1.50	1.06	4.96	4664.97	583.12	762
DDR-333 1GB Module	2.5	2.19	2.66	5.48	2057.06	257.13	245
DDRII-667 2GB Module	1.8	2.88	5.34	5.18	971.51	121.44	139
DDR3-1333 2GB Module	1.5	3.68	10.66	5.52	517.63	64.70	52
DDR4-2667 4GB Module	1.2	5.50	21.34	6.60	309.34	38.67	39
HMC, 4 DRAM w/ Logic	1.2	9.23	128.00	11.08	86.53	10.82	13.7

.087 W/GB/sec = .087 MW/PB/sec

Remember my BG/Q exascale strawman: 208 petabytes/sec?

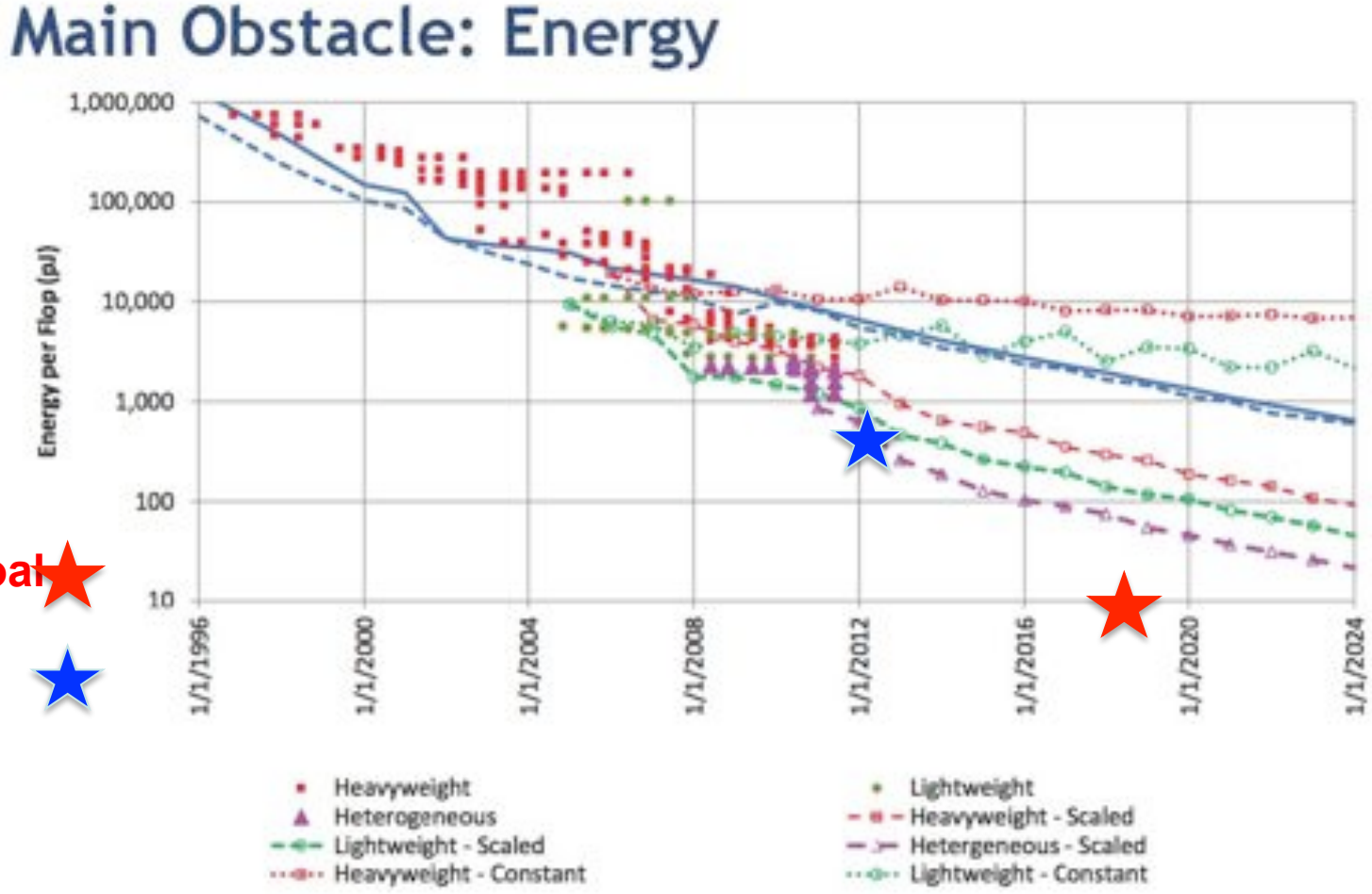
18 MW just for the memory system of such a machine!

But we might do 2x better by stacking 8 DRAMs, which is planned

So estimate 10 MW for the memory

Credit: J T Pawlowski of Micron @ Hot Chips 23, August, 2011

Energy (pJ) per flop



DOE goal

BG/Q

(courtesy P. Kogge)

Conventional architecture will NEVER make it!

Power summary table exascale at 11 nm in 2018

- Processors: 90 MW (1/3 scaling projection)
- Memory: 10 MW (based on HMC)
- Interconnect: 20 MW (fat tree overhead)
- DAV: 5 MW (5% of compute)
- Storage: 10 MW (10% of compute)
- Total: 135 MW/exaflops system
- 8 MW equates to ~60 PF
- If DoE's 20 MW/exaflops goal achieved ~400 PF is possible

NCAR

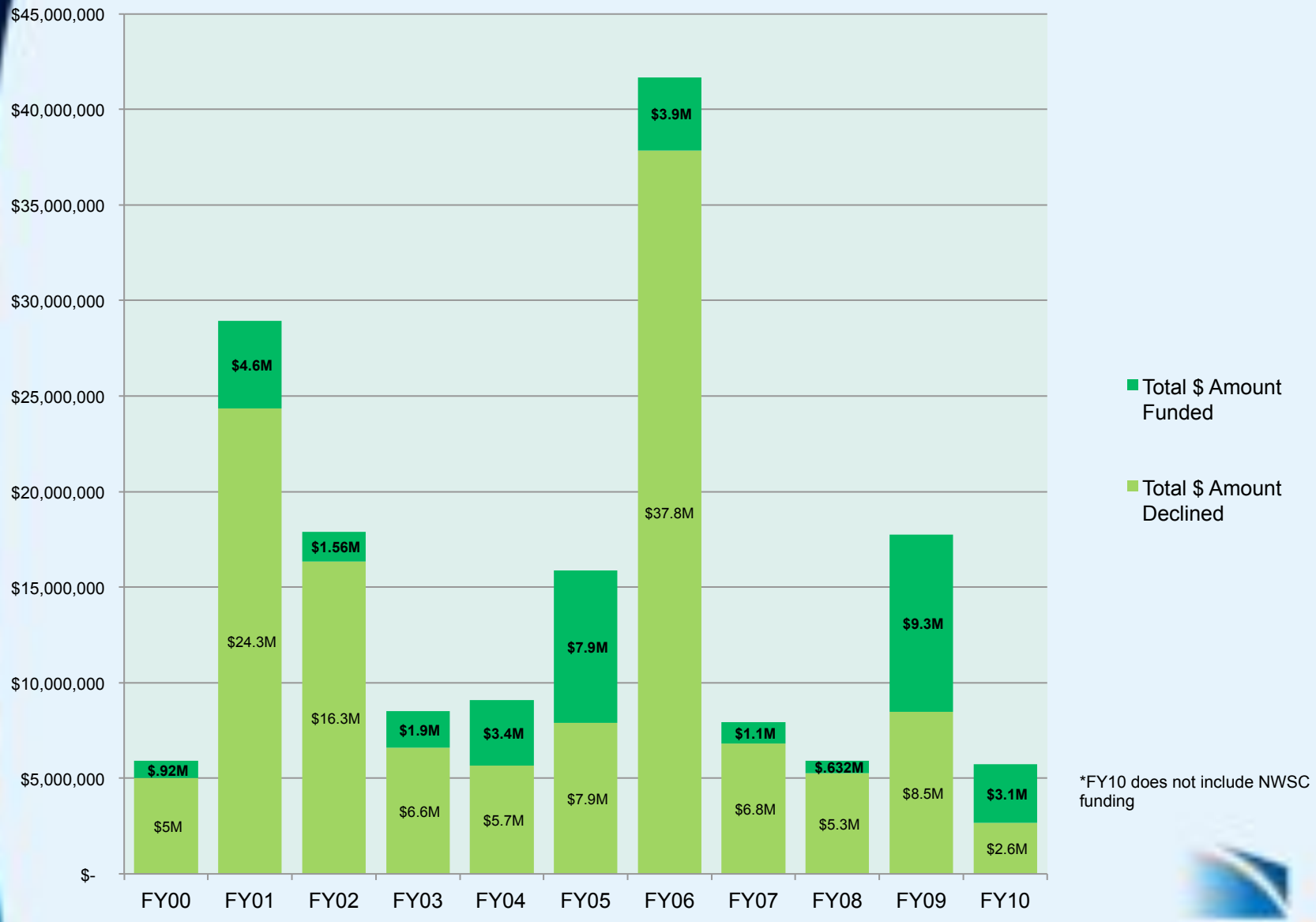
Questions?

NCAR

Proposal Status

NCAR

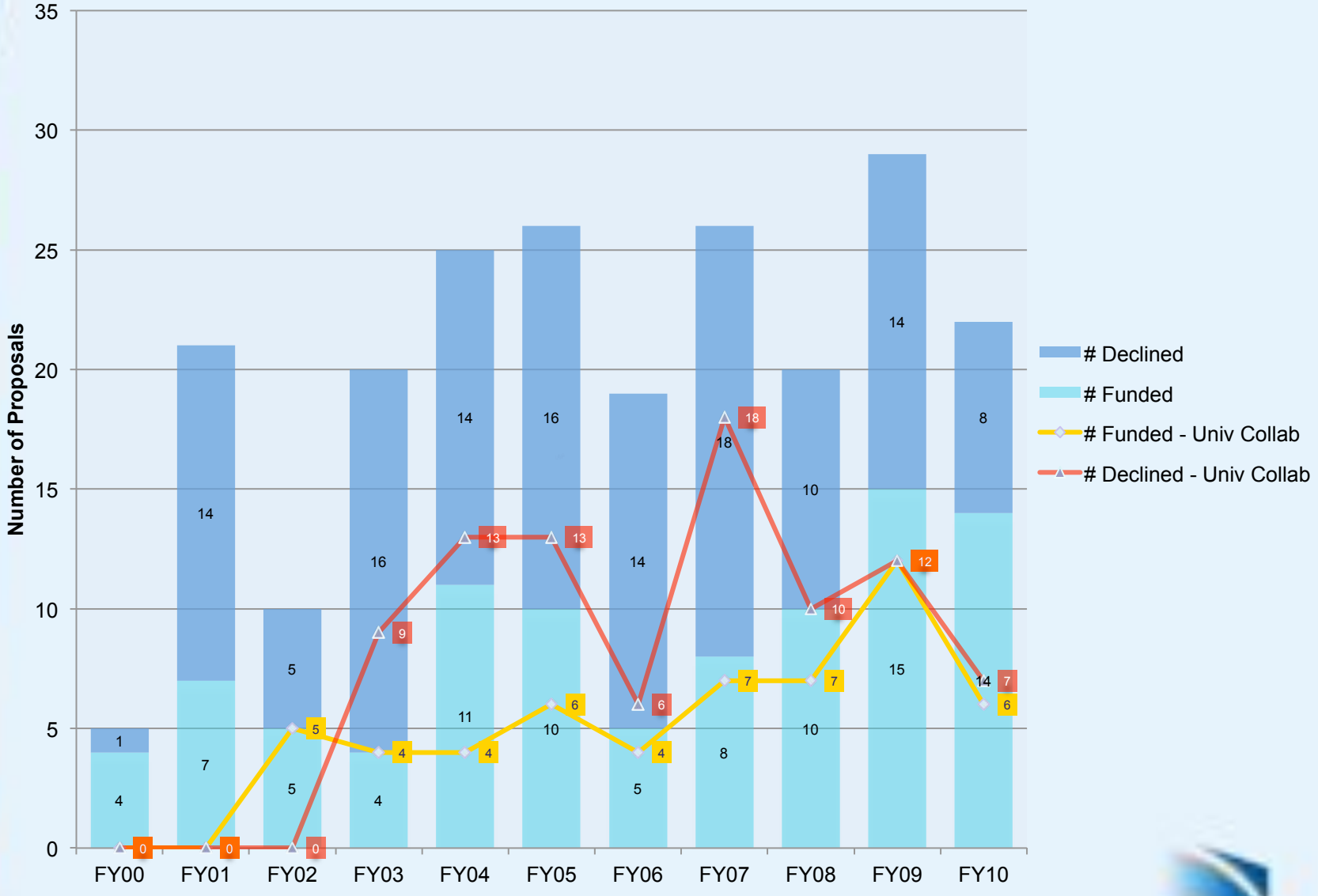
CISL Non-NSF Funding Success - $ Funded & Declined



*FY10 does not include NWSC funding

NCAR

CISL Non-NSF Funding Success - # Funded & Declined



NCAR

CISL Proposal Activity - 2011

Proposal Title	Funding Agency	Amount Requested	Length of Proposal	Status	CISL PI	Lead Submitting Institution
Modeling of Intra-Americas Sea Circulations to Assess Impacts on Climate Variability and Change over North America	NOAA	$96,362	3 years	Pending	Sain, Steve	UCSD/Scripps
Integration of multi-satellite altimetery data assimilation and hydrodynamic models for global discharge monitoring	NASA	$75,000	3 years	Pending	Anderson, Jeff	NASA
SI2-SSI: SciDaaS – Data management as a service for small/medium labs	NSF	$260,166	5 Years	Pending	Middleton	CISL
SI2-SSI CloudCentral: Providing and Outsourcing Model for Science	NSF	$195,000	3 Years	Pending	Woitaszek	U of Chicago
Compute and Storage Resources to support EaSM community	NSF	$1,847,177	1 Year	Funded	Kamrath	CISL/NCAR
Collaborative Research: SI2-SSI: A community ensemble data assimilation software facility for the geosciences	NSF	$84,358	5 Years	Pending	Anderson, Jeff	CISL/IMAGe/ NCAR
Computer support for the Antarctic Mesoscale Prediction System (AMPS) at the NCAR Computational and Information Systems Laboratory (CISL)	NSF	$700,000	1 Year	Funded	Kamrath	CISL/NCAR
RIVET: an expedition to develop a unifying community and middleware for pervasive and impactful immersive visualization	NSF	$0	5 Years	Pending	Clyne	Indiana University
Enabling regional climate model evaluation: A critical use of observations for establishing core NCA capabilities	NASA	$115,855	27 Months	Pending	Mearns	CISL/JPL

NCAR

CISL Proposal Activity - 2011

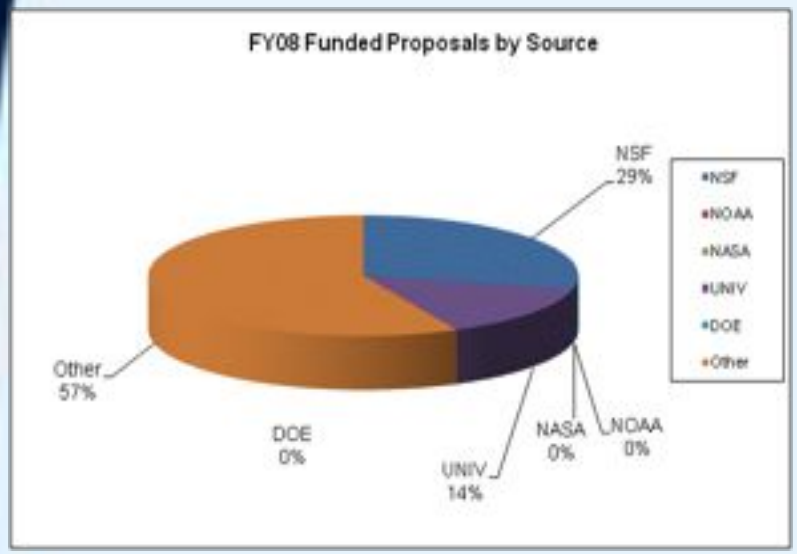
Proposal Title	Funding Agency	Amount Requested	Length of Proposal	Status	CISL PI	Lead Submitting Institution
Planning Letter: Office of Naval Research (ONR), Department Research Initiative (DRI)Emerging Dynamics of the Marginal Ice Zone	DOD	$0	5 Years	Pending	Middleton	EOL
Enhanced utilization of RO observations using advanced forward operators and error analysis in the ensembled data assimilation system with regional and global models	NASA	$429,487	3 Years	Pending	Anderson, Jeff	CISL/NCAR
High Productivity Computing Methods and Technology for High-resolution Earth Climate System Models.	NSF	$0	5 Years	Declined	Tufo	CISL/NCAR
Extreme value analysis as a tool for impact assessment of changes in extreme climate events on water quality	EPA	$128,496	3 Years	Declined	Katz	CISL/Miami
NWSC Transition Costs	NSF	$2,100,000	2 Years	Funded	Kamrath	CISL/NCAR
Multiple testing methods for random fields and high-dimensional dependent data	NIH	$70,789	4 Years	Pending	Sain	Harvard
SDCI Net: Collaborative Research: An integrated study of datacenter networking and 100 GigE wide-area networking in support of distributed scientific computing	NSF 10-504	$299,679	3 Years	Funded	Dennis	UVA
Collaborative Project:Ocean-Atmosphere Interaction From Meso- to Planetary Scale: Mechanisms, Parameterization, and Predictability	DOE	$479,815	3 Years	Funded	Dennis	CGD/NCAR
Collaborative Project: Interaction of Coastal and Estuarine Processes with Climate	DOE	$440,009	3 Years	Funded	Dennis	CGD/NCAR

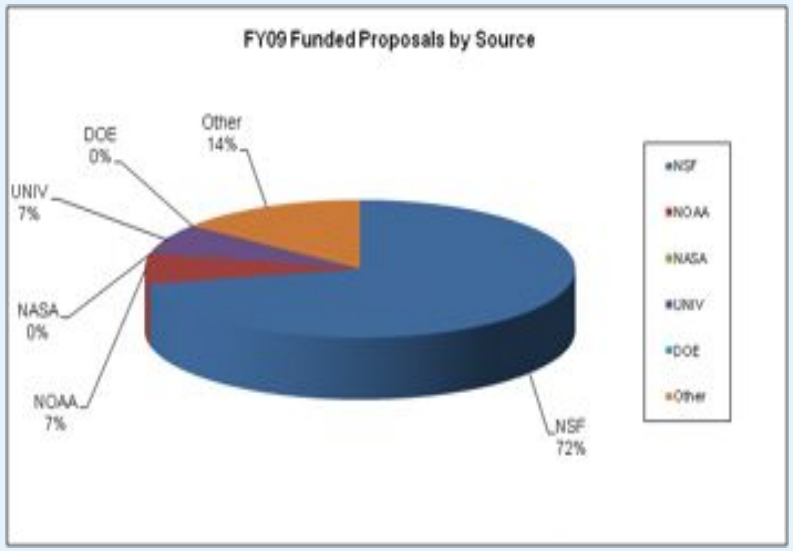
NCAR

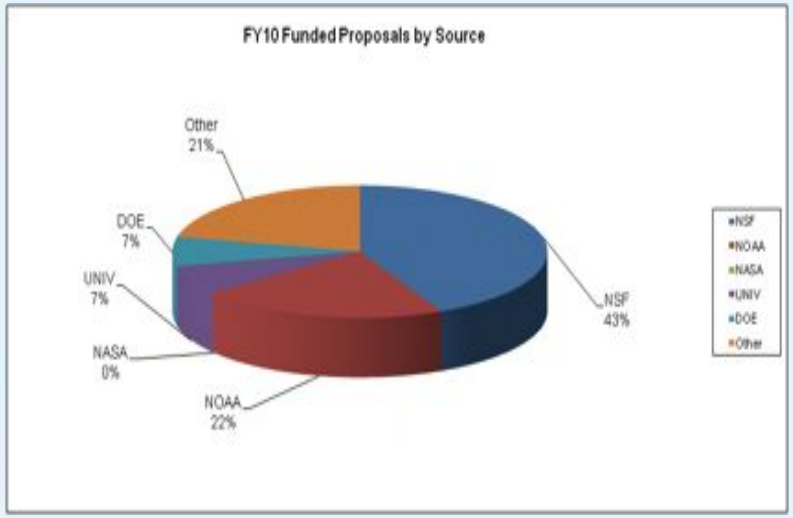
CISL Proposal Activity - 2011

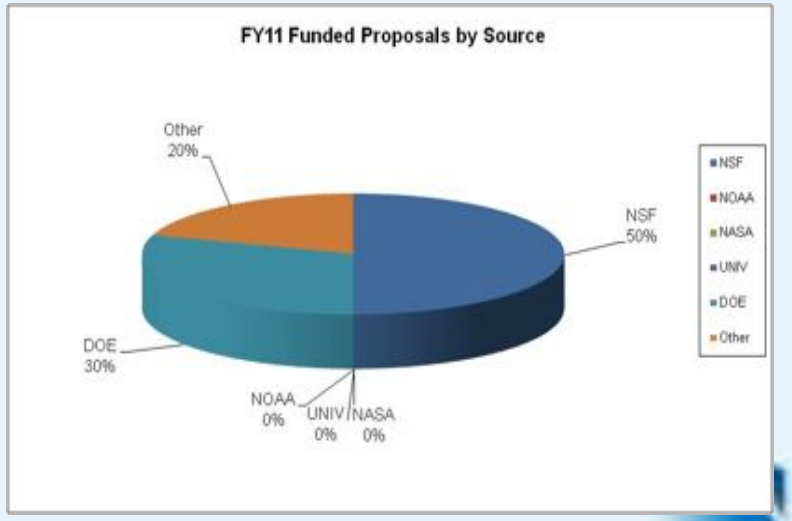
Proposal Title	Funding Agency	Amount Requested	Length of Proposal	Status	CISL PI	Lead Submitting Institution
Collaborative Project: Closing the Oceanic Branch of the Hydrological and Carbon Cycles and Sea Level Budget in CESM	DOE	$871,293	3 Years	Declined	Dennis	CGD/NCAR
Massive Datasets	NSF 10-592	$0	5 Years	Awarded	Nychka	UNC
Toward a non-hyrdrostatic High Order Method Modeling Environment (HOMME)	DOE	$459,826	3 Years	Awarded	Nair	CU-Boulder
Understanding data needs in decision making to manage vunerability to DoD installations: human and environmental decisions	DoD	$225,000	3 Years	Pending	Mearns	PNL
Holistic approach for assessment of climate-change related vulnerabilities at DoD installations: human and environmental dimensions	DoD	$300,000	2 Years	Declined	Mearns	Exponent
Decision Scaling: Tailoring climate information for DoD vulnerability assessment and adaption planning	DoD	$330,000	3 Years	Pending	Mearns	U Mass
FRGP Internet2 DYNES Network	Internet2	$0	1 Year	Awarded	Meehl	CISL/NCAR
From farm management to governance of landscapes; linkages and feedback between climate, ground water, land use and decisions and floods in the Argentine Pampas	NSF	$126,731	3 Years	Declined	Katz	CISL/NCAR
Visual data analysis tools for ocean model data	KISTI	$299,862	3 Years	Funded	Clyne	CISL/NCAR

FY08-FY11 CISL Funded Proposals by Fund Source









OTHER = NRL, DoD, Foreign, Commercial, etc.
(note: FY11 has 11 Pending Proposals)

NCAR