



Australian Government

Bureau of Meteorology

BUREAU OF METEOROLOGY



Efficiencies of Climate Computing an Australian perspective

Tim F. Pugh

Centre for Australian Weather and Climate Research

<http://www.cawcr.gov.au/>

Energy-Aware High Performance Computing

Hamburg, Germany

7th – 9th September 2011

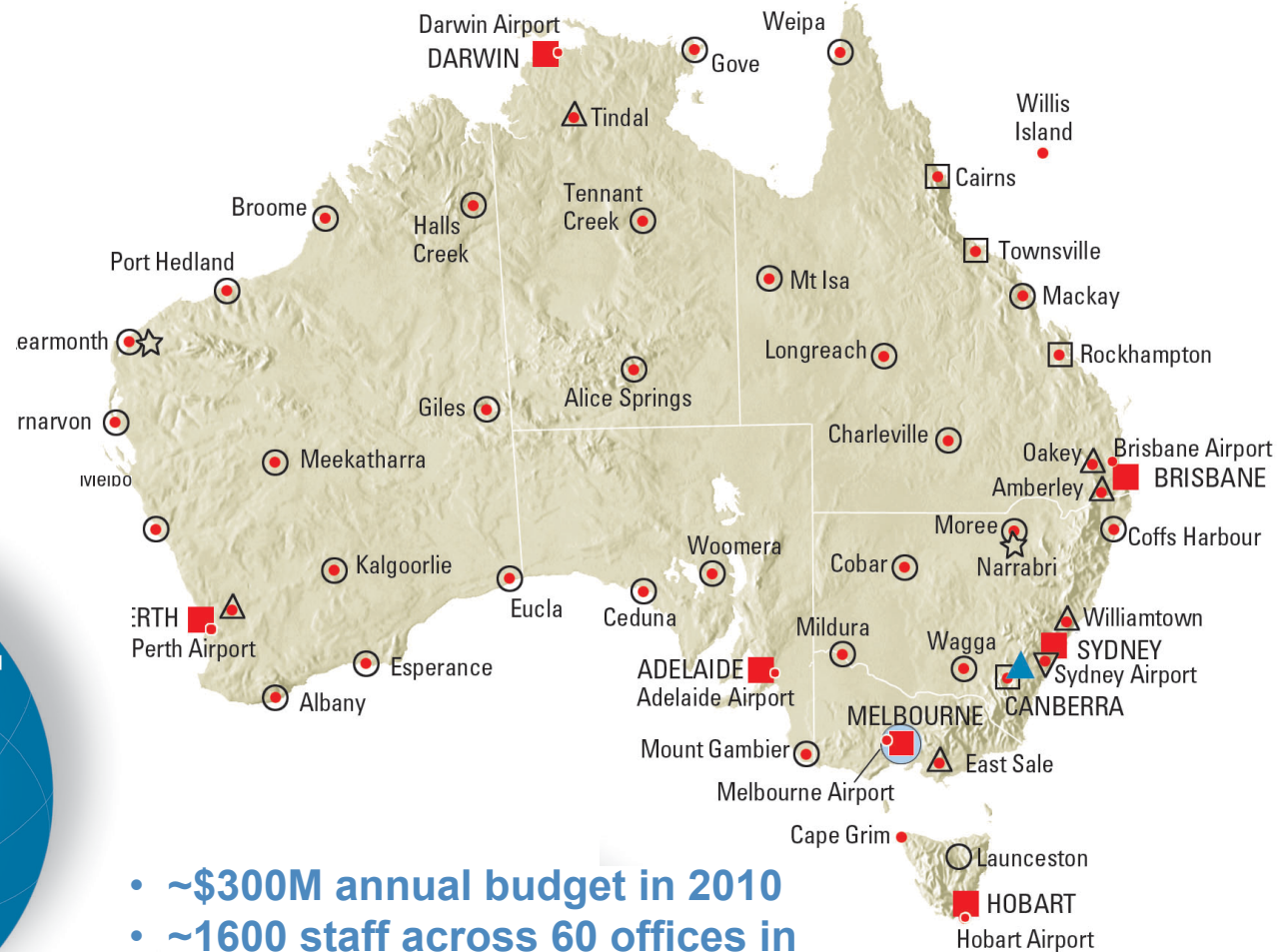
The Centre for Australian Weather and Climate Research
A partnership between CSIRO and the Bureau of Meteorology



Where the Bureau's staff work

LEGEND

- Regional Office
- Forecasting Office
- Information Office
- Airport Meteorological Unit
- ▲ Defence Meteorological Support Unit
- △ Defence Weather Service Office
- Staffed Observing Office
- ☆ Solar Observatory
- National Meteorological and Oceanographic Centre



- ~\$300M annual budget in 2010
- ~1600 staff across 60 offices in Australia, territories & Antarctica
- The Bureau's Head Office is in Melbourne



Australian Government
Bureau of Meteorology

Commonwealth Scientific and Industrial Research Organisation (CSIRO)

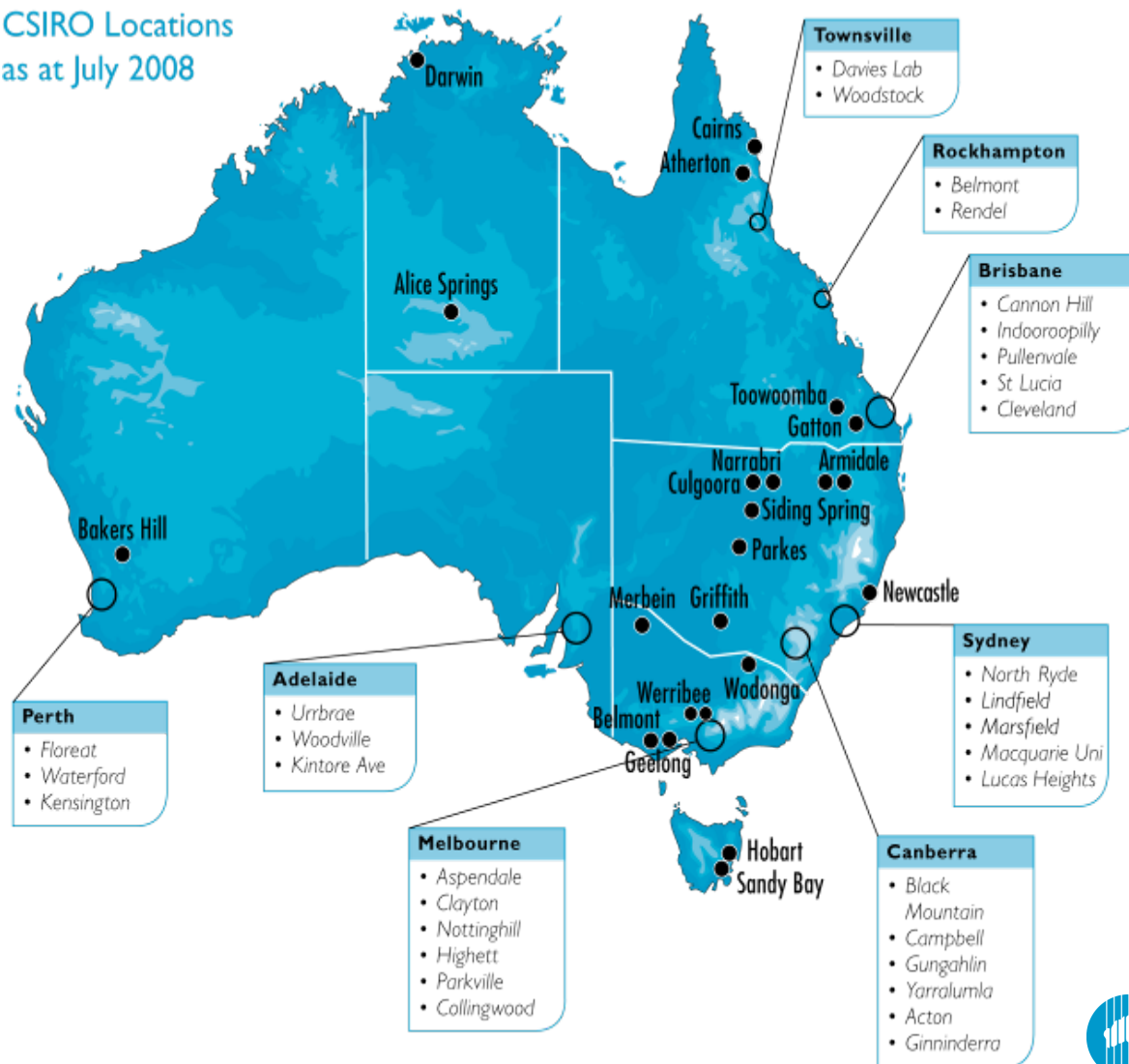


CSIRO

6400 staff located across more than 50 sites in Australia and overseas

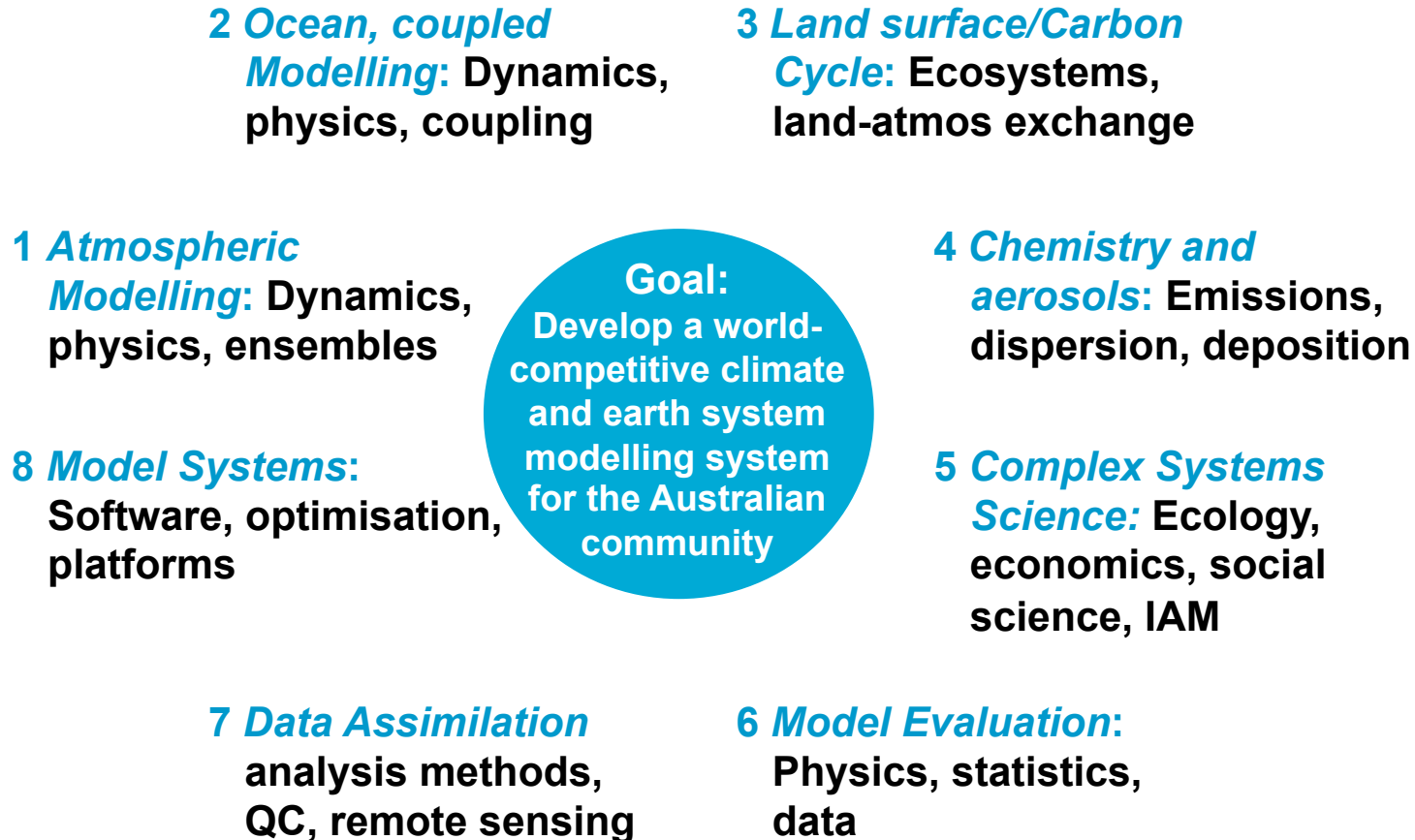
\$668 million budget in 2008-09

CSIRO Locations as at July 2008





Overview of capability – CAWCR Teams



Program members are well represented on international working groups



Australian models for CMIP5 simulations

CSIRO Mk3.6 – established coupled climate model

- Collaboration CAWCR/CSIRO and QCCCE
 - QCCCE (Queensland Climate Change Centre of Excellence)
 - <http://www.climatechange.qld.gov.au/>
- Utilises QCCCE machine for computation
- Utilises the NCI NF node of the Earth System Grid for output data distribution

ACCESS – new coupled earth system model

- Collaboration CAWCR (CSIRO and Bureau) and Universities
- Utilises NCI NF vayu machine for computation
- Utilises the NCI NF node of the Earth System Grid for output data distribution

Present TeraScale Systems

- Climate scientist have three Terascale systems available today
 - Bureau of Meteorology's HPC system called "Solar"
 - ANU/NCI HPC system called "Vayu"
 - CSIRO has a 24% share in the NCI HPC system
 - Climate modelling is primarily run at ANU/NCI
 - QCCCEE's SGI Altix ICE 8200 and InfiniteStorage procured in June 2010
 - 800 cores, 9.1 TF peak, 2.7 TB memory, 230 TB storage
- Bureau and NCI systems procured in 2009 from Sun Microsystems
 - Selection based on Performance and Energy Efficiency
 - Bureau imposed a 200 KW constraint on system designs

NCI, an initiative of the Australian Government, hosted by The Australian National University and is jointly funded by the Department of Innovation, Industry, Science and Research under its NCRIS program, CSIRO and ANU.

<http://nci.org.au/facilities-and-services/national-facility/>



Bureau's HPC System "Solar"

System Racks

6 x C48 blade racks
5 x 19" racks
200kW power (typical)
220kW power (HPL)
49 Tflops HPL
Water and Refrig cooling

576 Compute nodes

4,608 Cores
~54 Peak Tflops
13.8 TiB memory
24GB SSD drives
QDR 4x IB optical network

Lustre Parallel File System

18 x OSS nodes
18 x J4400 JBOD
115 TB usable storage
8 GBps Bandwidth



Sun Constellation System
Commissioned in June 2010

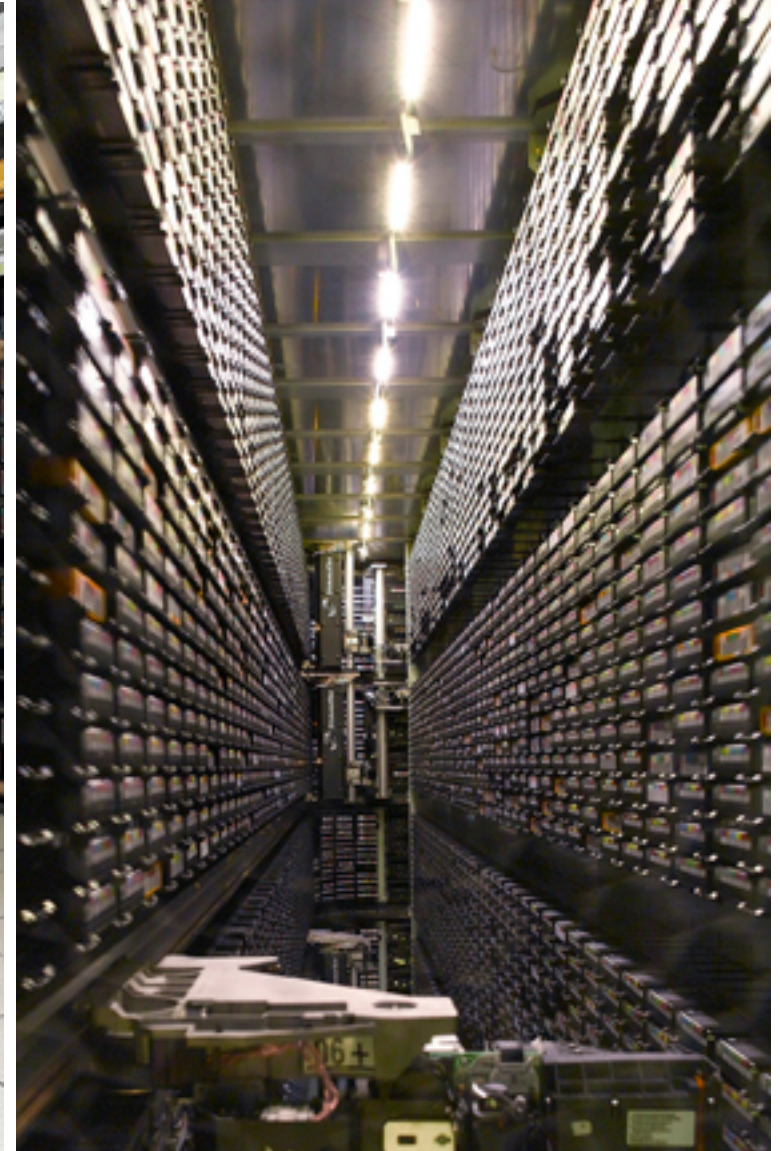


Australian Government
Bureau of Meteorology

Large Scale Data Storage System



2.5 PB (10,000 tape) SL8500 tape archive
With Sam-FS, TSM (MARS)
T10000 and LTO4,5 tapes





LSDSS Systems

- **Storage Tiers**
 - HDS 9990v for mission critical storage
 - Tier 1 – 180 TB mission critical disks (FC disks)
 - Tier 2 – 260 TB main storage (SAS disks)
 - Tier 3 – 2000 TB bulk storage (SATA disk)
 - Sun 6540, 6580's FC disk arrays
 - Tier 4 – 300 TB Virtual Tape Library
 - Tier 5 – 2.5 PB (10,000 tape) SL8500 tape archive
 - With Sam-FS, TSM (MARS)
 - T10000 and LTO4,5 tapes
- **HSM storage mgmt system**
- **ESM SAN fabric mgmt**
- **NAS File Service – pairs of BlueArc 2100 & 3200**
- **HPC Storage (SAS) – Lustre File System on Solar**
 - Data volumes are presently growing at ~50% per annum



Australian Government
Bureau of Meteorology

ANU/NCI HPC System “vayu”

System Racks

16 x C48 blade racks
6 x 19” racks
34 sq m floor space
605 kW power (Typical)
750kW power (HPL)
127 Tflops HPL

1,492 Compute nodes

11,936 Intel Nehalem cores
140 Peak Tflops
36 TiB memory
24 GB SSD drives
QDR 4x IB optical network

13 Lustre Object Store pairs

26 x OSS nodes
52 x J4400 JBOD
834 TB usable storage
25 GBps BW



Sun Constellation System

Commissioned in April 2010

<http://nf.nci.org.au/facilities/vayu/hardware.php>



NCI Storage Systems

- NCI recently completed a Storage tender
 - SGI won the contract
 - 8 racks x 700 TB usable= 5.6 PB usable storage
 - three tape racks and 2200 (1.5TB) tapes - Spectra Logic T950
 - Each rack frame can scale out to 15PB
- HPC Storage – Lustre 1.8.6 File System on Vayu
 - 834 TB usable SATA storage
 - 25 GBps bandwidth
- Additional storage capacity expected
 - RDSI program (\$50M) for research storage infrastructure
 - Next Petascale HPC system in 2012



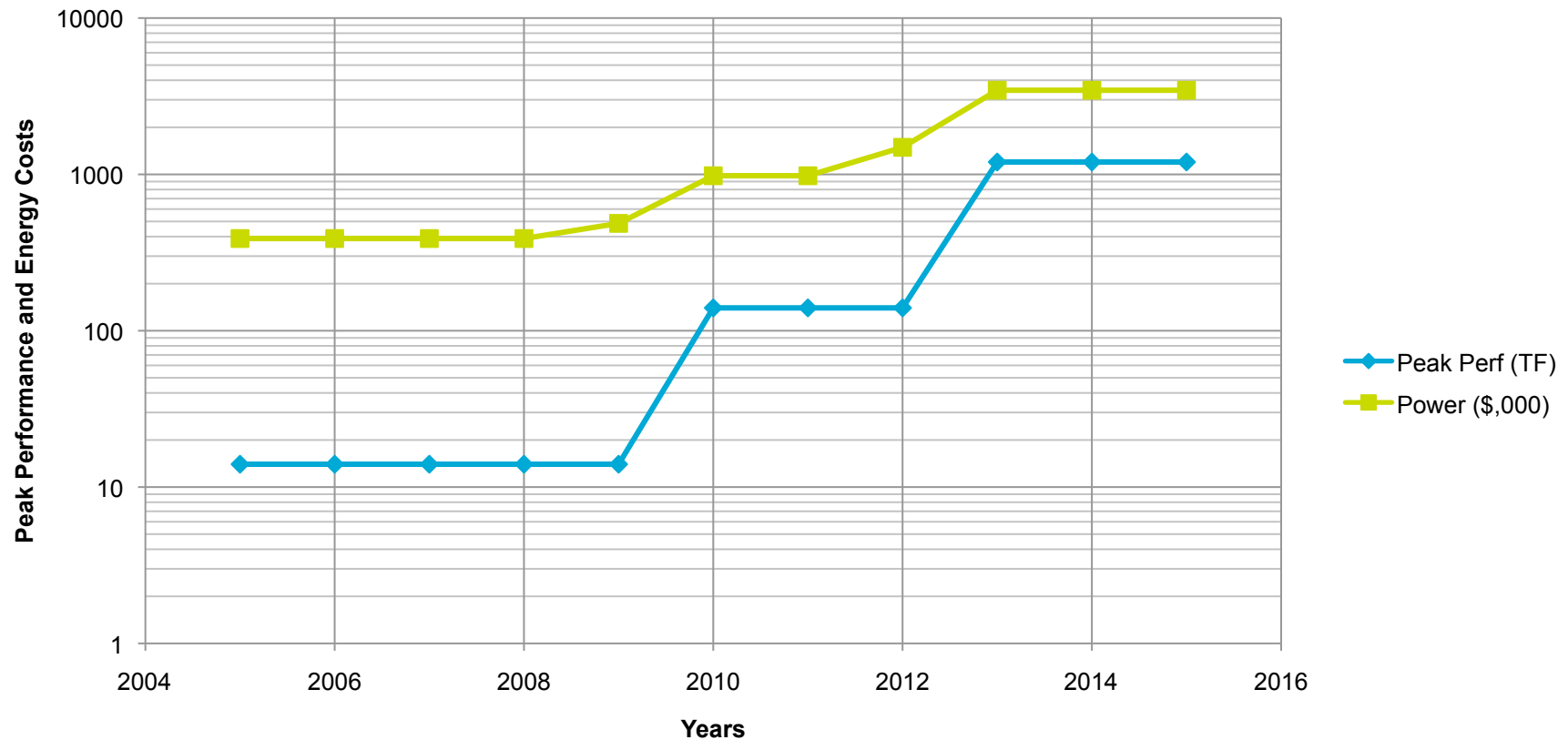
Energy Costs at NCI

- NCI facility has ...
 - up to 1.5 MW electrical power available
 - less than 700 KW used for computers, storage, cooling, and building
 - Electrical power costs are approximately \$750k p.a.
- Annual budget for electrical power is expected to significantly increase, and grow with the new Petascale system in 2012
 - Future system procurements could be constrained by operating costs.
 - Electricity is expected to rise from \$0.125 to \$0.20 per kw-hr
 - New data centre is designed to be energy efficient, PUE < 1.2
 - Provides some offset from rising costs
 - Energy efficiency gains from ..
 - 100% heat removal with water cooled compute rack doors,
 - free-cooling due to higher water temperatures,
 - higher ambient air temperatures in data centre rooms, and
 - efficient electrical systems to avoid losses.



Energy Costs Estimates

Energy Costs History at NCI



Note: Years beyond 2011 are forward looking estimates



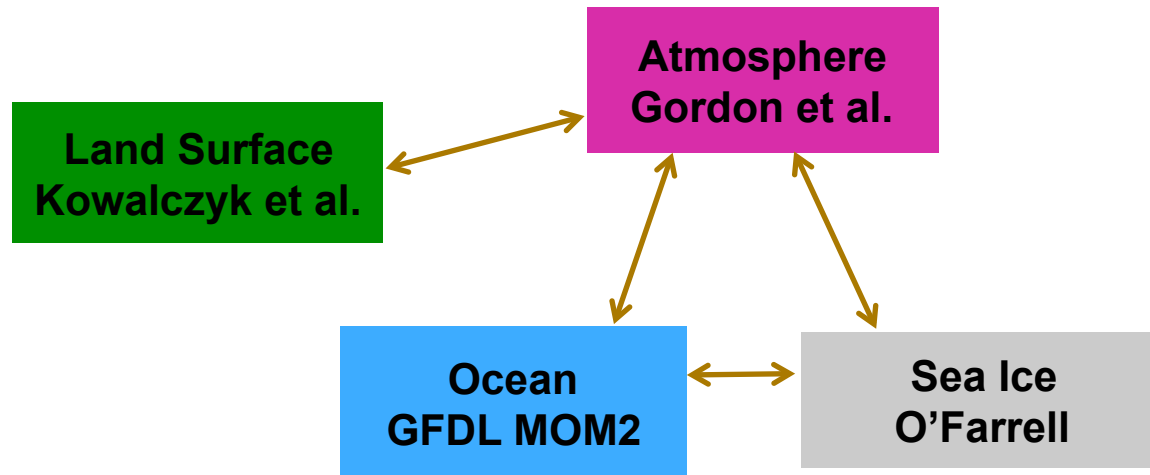
Australian Government
Bureau of Meteorology

2012 PetaScale HPC Facility

- April 2010, the Australian government announced \$50m contract to ANU/NCI to build a new Petascale HPC facility for climate change, earth system science and national water management research
 - To procure a new data centre
 - To procure a petascale computing system
 - Partners to provide operating budget of facility
 - Target date for system delivery is 2H 2012
- <http://nci.org.au/news-and-events/news/funding-agreement-signed-a-a-new-petascale-high-performance-com/>
- National Computational Infrastructure (NCI) is located at the Australian National University (ANU), Canberra, Australia



Mk3.6 coupled climate model for AR5



- **Atmosphere and sea ice:** T63 – 1.875° lon x 1.875° lat; 18 levels
- **Ocean model:** 1.875° lon x 0.938° lat, enhanced tropical ; 38 levels



Australian Government
Bureau of Meteorology

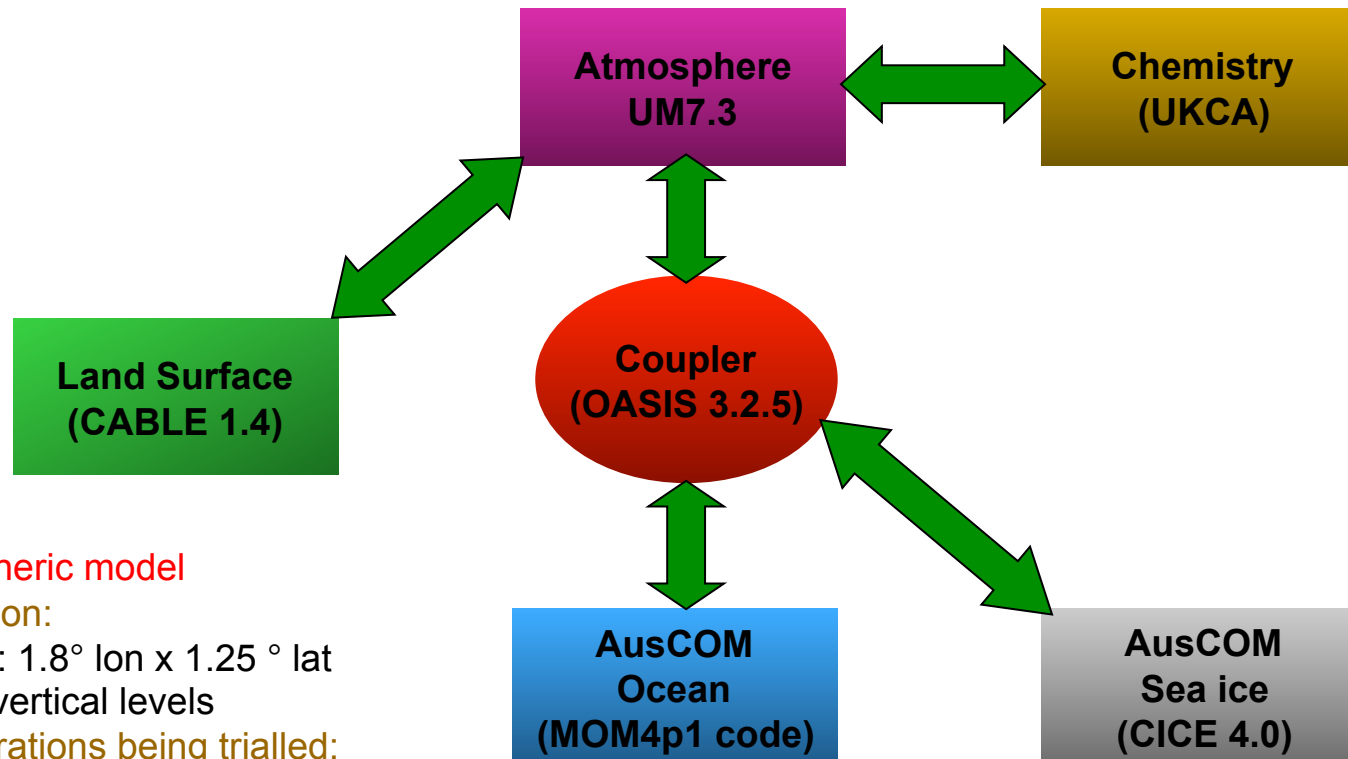
CSIRO Mk3.6 – CMIP5 simulations

Experiment	Length	Ens.
Control	500 yr (160 yr spin up)	1
Historical	1850-2005	10
AMIP	1970-2008 (15 yr spin up)	10
Mid-Holocene	100 yr (300 yr spin up)	1
RCPs 2.6, 4.5, 8.5, 6.0	2006-2100 (<i>RCP 4.5: 3x 2100-2300</i>)	10
1%/yr CO ₂ to 4x	140	1
AGCM + control SSTs	30	1
AGCM + control SSTs + 4x CO ₂	30	1
4x CO ₂	150 + 5	1+11
AGCM + control SSTs + AA	30	1
AGCM + control SSTs + SA	30	1
Historical (natural)	1850-2005	10
Historical (GHGs)	1850-2005	10
Historical (anthropogenic)	1850-2005	10
Historical (all except ozone)	1950-2005	10
Historical (all except AA)	1850-2005	10
Historical (AA)	1850-2005	10
Historical (Asian aerosols)	1850-2005	10





ACCESS CMIP5 Modelling System



Atmospheric model

Resolution:

N96L38: 1.8° lon x 1.25° lat
and 38 vertical levels

Configurations being trialled:

HadGEM2

HadGEM2 + PC2 clouds

Proto-HadGEM3

3 hourly flux coupling between models

3.5 simulated years / day

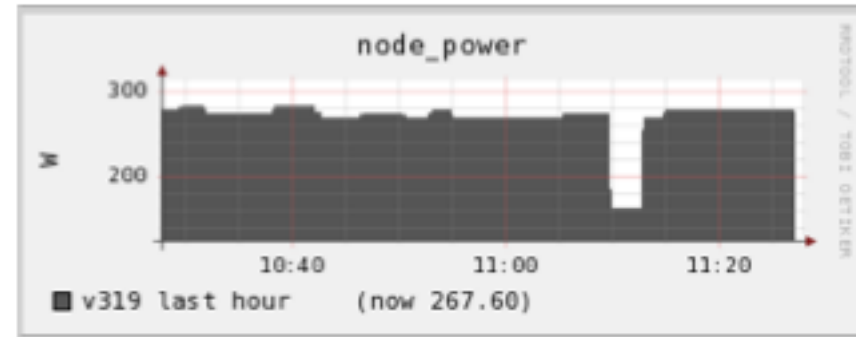
**CABLE, AusCOM, CICE,
UKCA have been
successfully coupled to
Unified Model**



Climate Computing in 2011

Climate Computing in 2011

# cores per node	8
Simulated years per Day	3.5
Watts per compute node	270
System Watts per compute node	405
Operating (\$ per kilowatt-hr)	\$0.13
System TCO (\$ per cpu-hr)	\$0.13



Component	Atmosphere N96L38	Ocean 1 degree res	Sea Ice 1 degree	OASIS 1 MPI task	System
Configuration					
# of Cores	96	40	6	1	144
Kilowatts per simulated year	33.4	13.9	2.1	0.3	50.0
\$Cost(power) per simulated year	\$4.17	\$1.74	\$0.26	\$0.04	\$6.26
\$Cost(asset) per simulated year	\$82.29	\$34.29	\$5.14	\$0.86	\$123.43
Total Costs per simulated Year	\$86.46	\$36.02	\$5.40	\$0.90	\$129.68
% of Total	66.67%	27.78%	4.17%	0.69%	

A 100-year simulation costs \$12,968 and consume 5,000 kw-hrs. Recent model runs are achieving 4-5 simulated years per day.



ACCESS – CMIP5 “Core” simulations (long term) in 2011, for AR5

Experiment	Length	Ens.
Preindustrial Control	500 yr (~200 yr spin up)	1
Historical	1850-2005	1
AMIP	1979 (or earlier?) - 2008	1
RCPs 4.5, 8.5	2006-2100	1
1%/yr CO ₂ to 4x	140	1
AGCM + control SSTs	30	1
AGCM + control SSTs + 4x CO ₂	30	1
4x CO ₂	150	1

AR5 total costs estimated at \$146,673 and consumes 56,606 kw-hrs
Actual costs, multiple by the number of unsuccessful runs

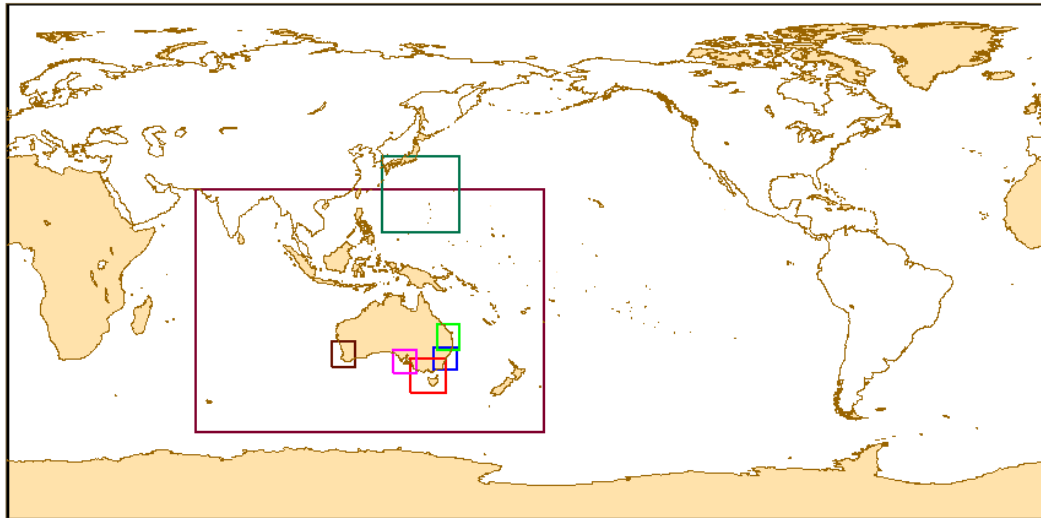


Energy Costs for Climate Computing

- AR5 climate modelling for 5th IPCC Assessments
 - Today Australia's climate computing has up to 25 M core hours available
 - NCI offers about 104M core hours per year today, and the new Petascale machine will hopefully offer 10x that amount.
 - Total costs of ACCESS's AR5 experiments are ~\$146k, \$7k for electricity.
 - Energy costs for CMIP5 will be significantly larger.
- TCO of NCI per year: approximately AU\$11 M
- Processor hours per year: approximately 400 M
- Prize per processor hour: about 12.5 cents
- These estimated costs are good relative to other climate sites!
 - However what is the intangible values at each site for the service and research support, the staff skills, visualization, data management, and computational assistance for projects.



Today's Weather is Tomorrow's Climate



“APS1”: ACCESS Parallel Suite 1

Met Office Unified Model vn7.5

Met Office 4D-VAR v26.1

- G – Global (40km L70)
- R – Regional (12km L70)
- C – City (4km L70)
- TC – Tropical Cyclone (12km L70)

NWP System	Domain	Type	Q1 2011	Q4 2012	2013	2015	2016	2017
			APS1	APS2	APS3	APS4	APS5	APS6
ACCESS-G	Global	10 day FCST	N320L70 (40 km) 640x481x70	N512L70 (25 km) 1024x769x70	20kmL90 (20 km) 1280x961x90	20kmL120 (20 km) 1280x961x120	15kmL120 (15 km) 1728x1297x120	10kmL120 (10 km) 2560x1921x120
ACCESS-R	Australian Region	3 day FCST	12kmL70 1090x750x70	12kmL70 1090x750x70	10kmL90 1200x825x90	10kmL120 1200x825x120	8kmL120 1500x1030x120	5kmL120 2400x1650x120
ACCESS-C	Cities	2 day FCST	5kmL70 160x160 to 240x240	2kmL70 400x400 to 600x600	1.5kmL70 533x533 to 800x800	1kmL70 800x800 to 1200x1200	1kmL90 800x800 to 1200x1200	1kmL120 800x800 to 1200x1200
ACCESS-TC	TC & Severe Wx	3 day FCST	12kmL70 300x300x70	12kmL70 300x300x70	10kmL90 330x330x90	8kmL90 450x450x90	8kmL120 450x450x120	4kmL120 900x900x120





Computational Needs for greater resolution

NWP System	Domain	Type	Q3 2009	Q1 2011	Q4 2012	2013	2015	2016	2017
			APS0	APS1	APS2	APS3	APS4	APS5	APS6
			Intel Nehalem	Intel Nehalem	Intel Sandy Bridge	Intel Sandy Bridge	2014 Processor	2014 Processor	2016 Processor
ACCESS-G	Global	10 day FCST 2 FCST / day	N144L50 (measured)	N320L70 (measured)	N512L70 (measured Nehalem)	20kmL90 (estimated)	20kmL120 (estimated)	15kmL120 (estimated)	10kmL120 (estimated)
		s-factor	1.00	6.90	2.56	2.01	1.33	1.82	2.19
		grid pts	3,124,800	21,548,800	55,121,920	110,707,200	147,609,600	268,945,920	590,131,200
		t-factor	1.00	1.00	1.60	1.25	1.00	1.35	1.48
		timestep (min)	15.00	15.00	10.00	8.00	8.00	5.93	4.00
		cores	240	640	782	1964	1309	3220	4652
		% System (cores)	5.21%	13.89%	8.49%	21.31%	4.74%	11.65%	12.62%
		elapse time (min)	35.00	81.17	84.83	84.83	84.83	84.83	84.83
		node-hr/ FCday	2	11	7	17	8	19	21
		cpu-hrs/FCST	140	866	1106	2777	1851	4553	6578
		cpu-hrs/day	280	1732	2212	5553	3702	9106	13156
		% System (cpu-hrs)	0.25%	1.57%	1.00%	2.51%	0.56%	1.37%	1.49%
		Gflops (peak)	2640	7040	14080	35348	47131	115928	376849
		Gflops (sustained)	132	352	704	1767	2357	5796	18842

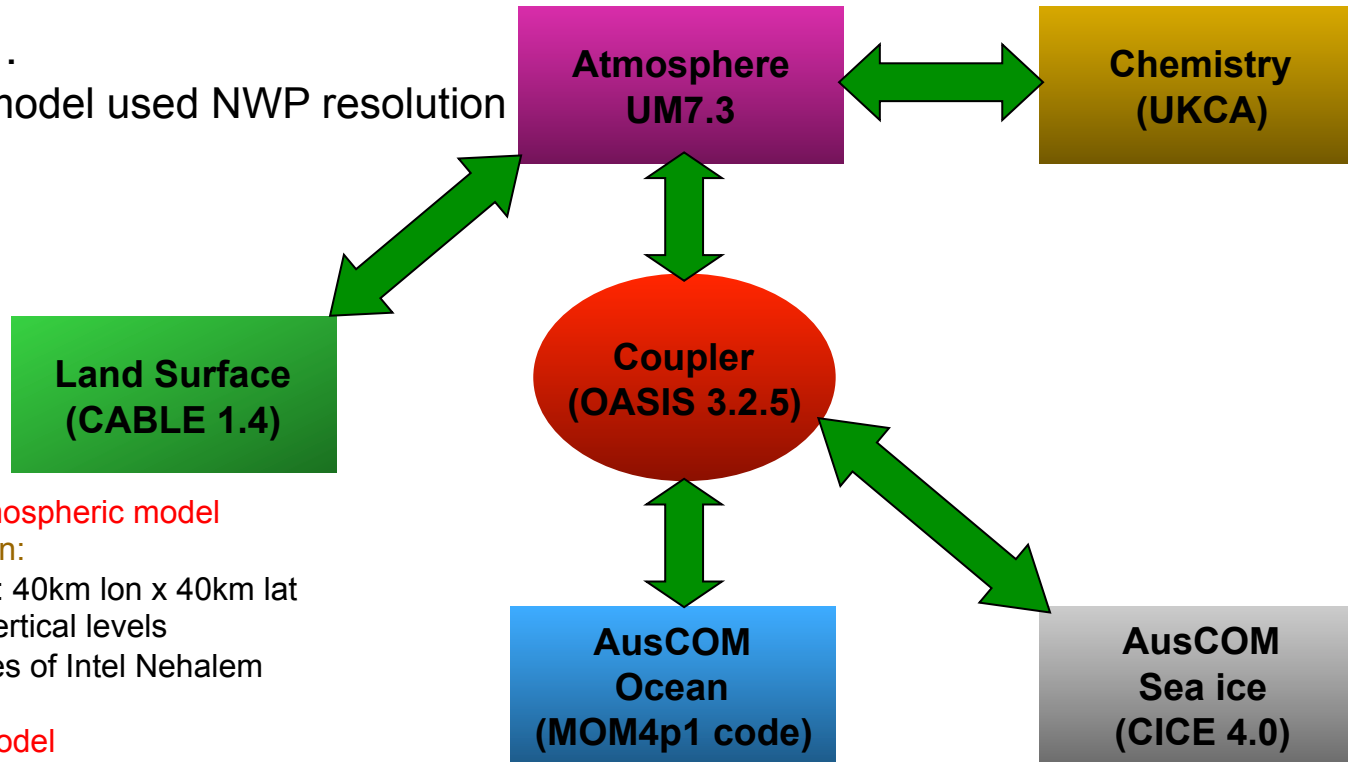
ACCESS-G Storage from 2009 to 2017 is 188 times increase



Climate Computing at higher resolutions

What if...

Climate model used NWP resolution



NWP Atmospheric model

Resolution:

N320L70: 40km lon x 40km lat

And 70 vertical levels

1120 cores of Intel Nehalem

Ocean model

Resolution:

¼ degree: 19km lon x 19km lat

And 47 vertical levels

232 cores of Intel Nehalem

3 hourly flux coupling between models

3.5 simulated years / day



Hi-Res Climate Computing in 2011

Climate Computing - Intel Nehalem

# cores per node	8
Simulated years per Day	3.5
Watts per node	270
System Watts per compute node	405
Operating (\$ per kilowatt-hr)	\$0.13
Asset (\$ per cpu-hr)	\$0.13

What if...increase climate resolution to current NWP resolutions to improve projections, fixing the number of simulations per day.

Component	Atmosphere	Ocean	Sea Ice	OASIS	System
Configuration	N320L70	1/4 degree res	1/4 degree	MPI tasks	
# of Cores	1120	232	28	4	1384
Kilowatts per simulated year	259.2	53.7	6.5	0.9	320.3
\$Cost(power) per simulated year	\$32.40	\$6.71	\$0.81	\$0.12	\$40.04
\$Cost(asset) per simulated year	\$960.00	\$198.86	\$24.00	\$3.43	\$1,186.29
Total Costs per simulated Year	\$992.40	\$205.57	\$24.81	\$3.54	\$1,226.32
% of Total	80.92%	16.76%	2.02%	0.29%	

A 100-year simulation costs \$122,632 and consume 32,030 kw-hrs. (6x increase in energy usage, and 9.4x increase in total costs)



ACCESS – CMIP5 “Core” simulations for Hi-Res Climate Computing



Experiment	Length	Ens.
Preindustrial Control	500 yr (~200 yr spin up)	1
Historical	1850-2005	1
AMIP	1979 (or earlier?) - 2008	1
RCPs 4.5, 8.5	2006-2100	1
1%/yr CO ₂ to 4x	140	1
AGCM + control SSTs	30	1
AGCM + control SSTs + 4x CO ₂	30	1
4x CO ₂	150	1

AR5 total costs estimated at \$1,386,971 and consumes 362,256 kw-hrs

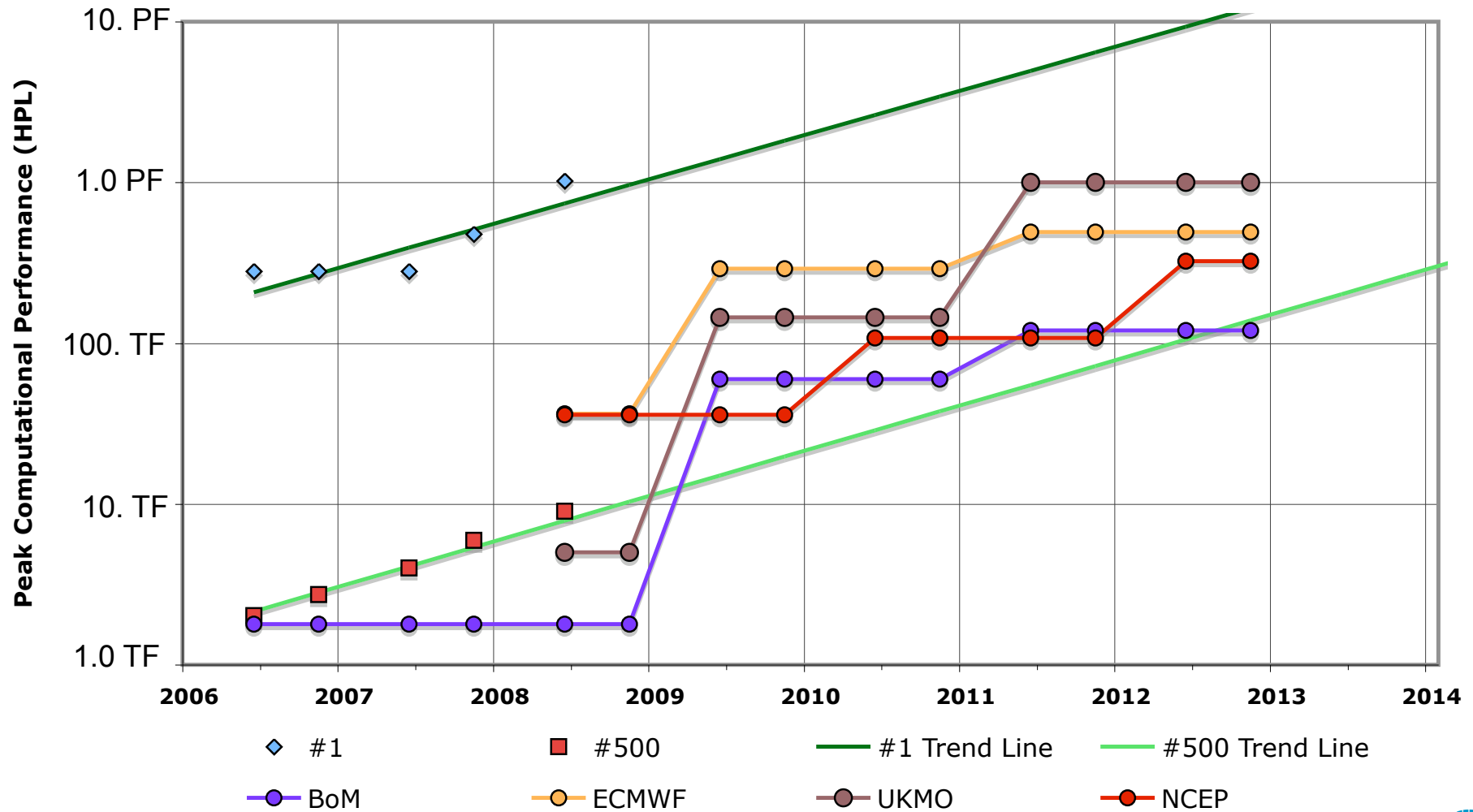
Efficiency of Climate Modelling

- **Increasing climate grid resolution (Capability computing)**
 - Drives cores counts to weather forecasting cores counts (scalability issues)
 - To maintain throughput (simulation years per day), climate modelling will equal or exceed NWP models core counts (HiRes climate > NWP)
 - High resolution climate modelling is capability computing with long time integrations!
 - 100 year run with 3.5 sim years/day is one month of computing
 - Lots can issues can arise on a system in one month!
- **Increasing model components (Capacity computing)**
 - Linear multiplier of memory and computing requirements
 - Increases in latency of flux data exchanges can slow throughput
- **Ensemble climate modelling (10-24x capacity computing)**
 - Linear multiplier of memory and computing requirements
 - Models need to run concurrently due to the length of the run.
- **Climate modelling is as challenging as NWP if not more.**



Supercomputing Projections

Projected Performance with respect to Top 500

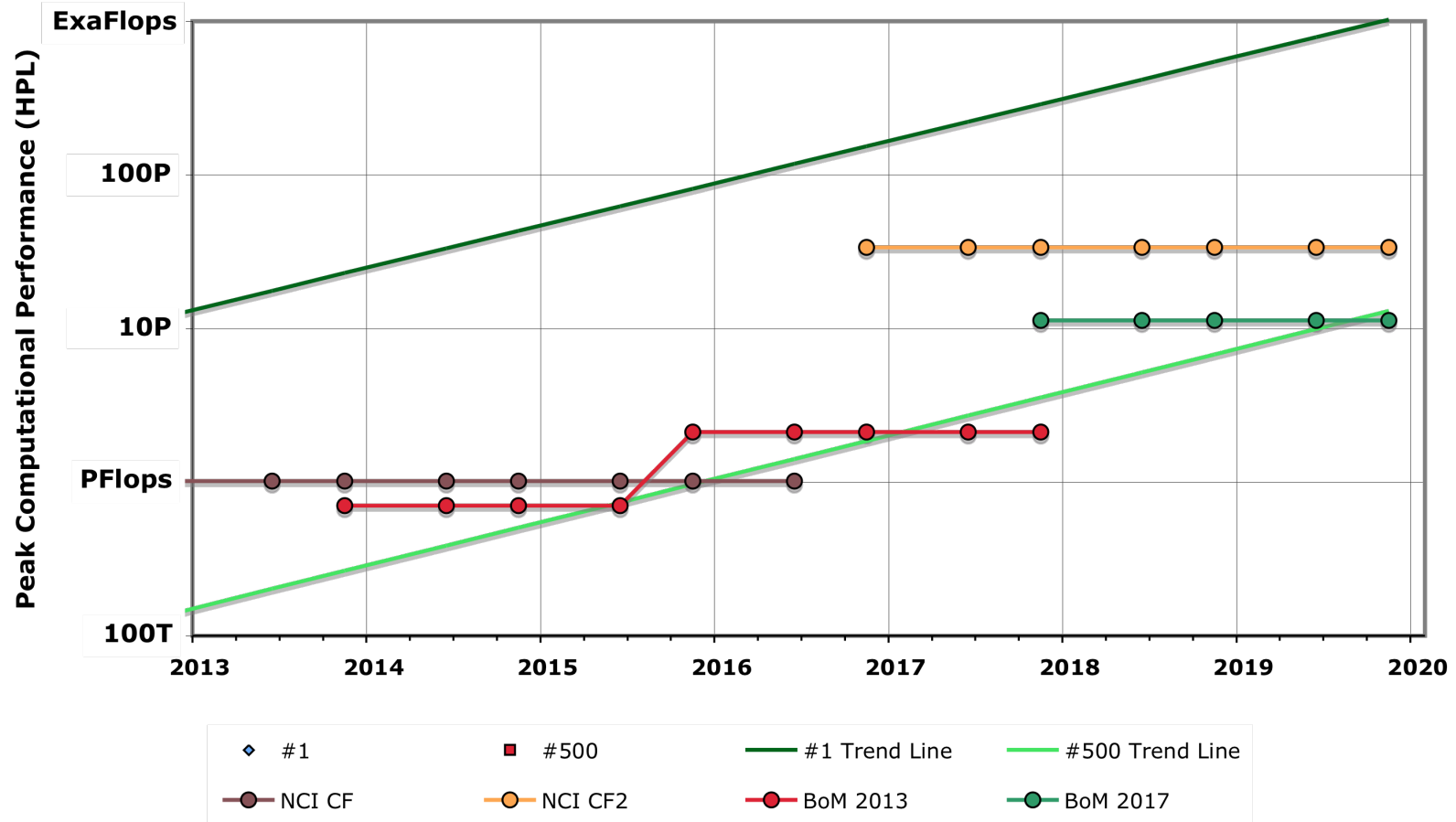


Source: 2008 WNGE Met centre projections/estimates



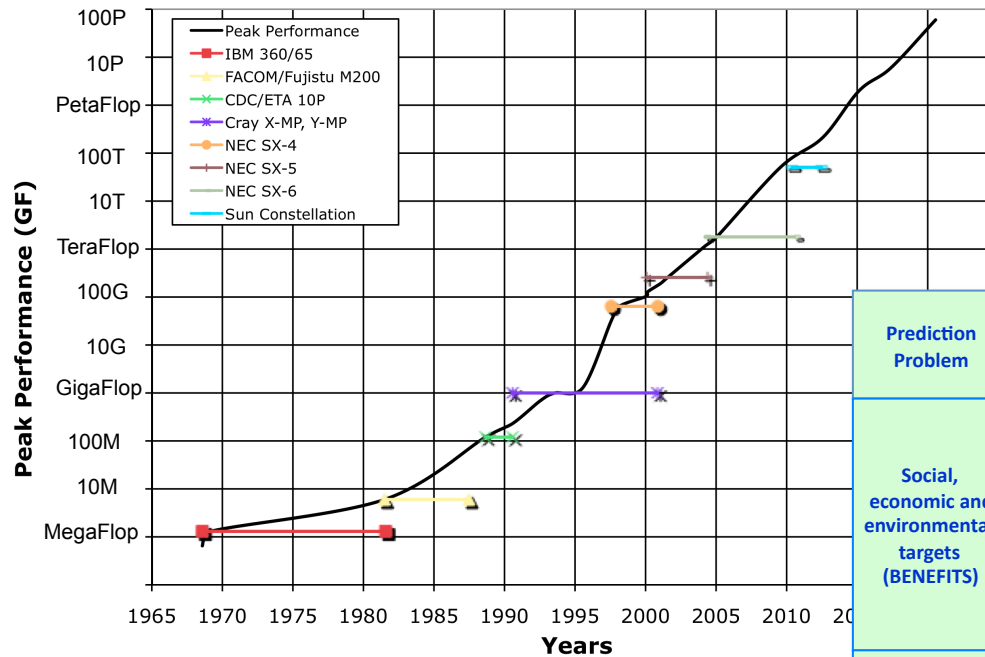
Estimate of HPC Tomorrow

Estimate of HPC Tomorrow





Seamless Prediction



Future developments will be underpinned by rapidly increasing computer power

Australia needs to grow supercomputing capacity to achieve insight and foresight into Climate and Environmental processes and prospects.

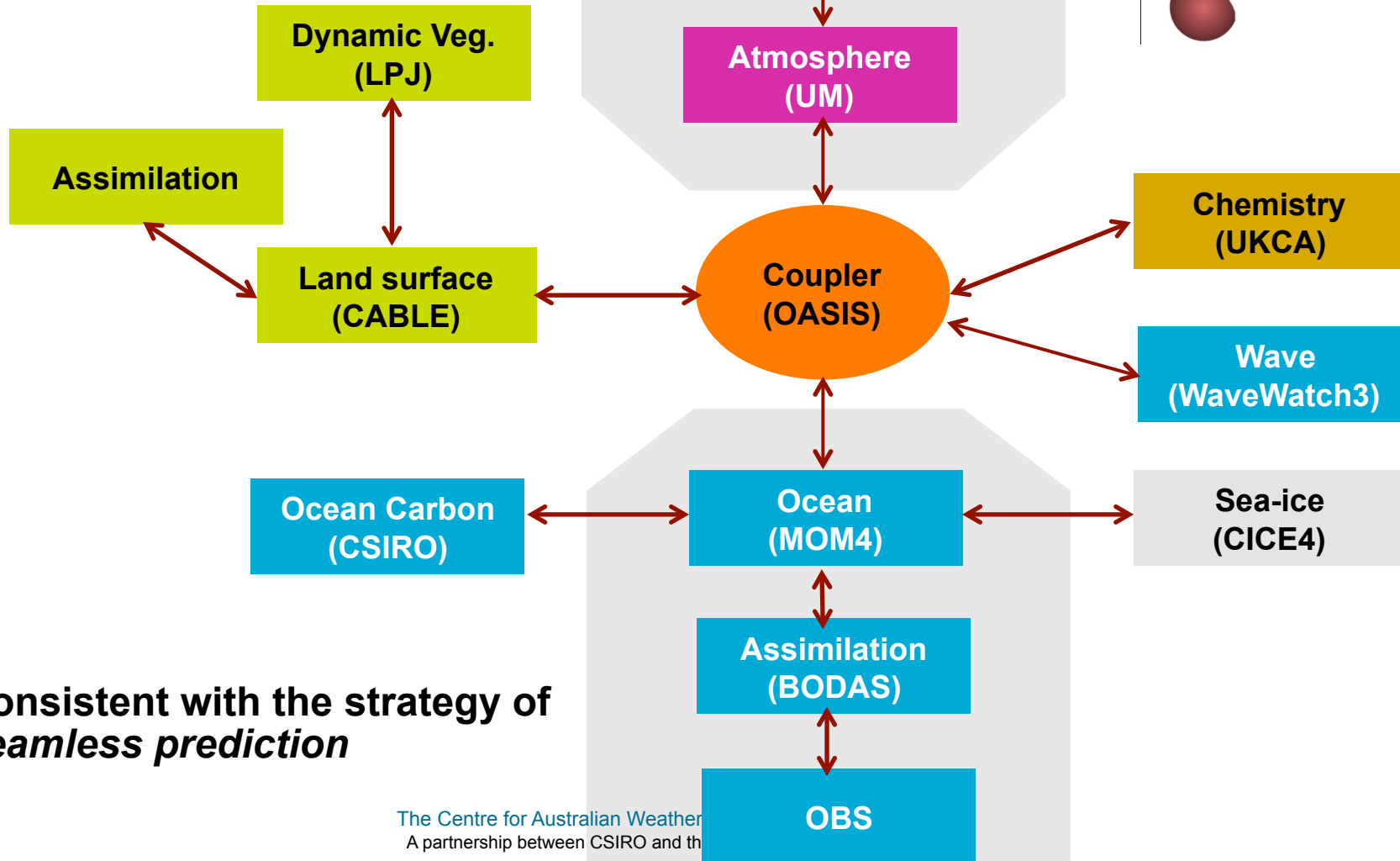
Prediction Problem	Nowcasting	Numerical Weather Prediction	Intra-seasonal Prediction	Seasonal Prediction	Decadal Prediction	Climate Change Projections	
Social, economic and environmental targets (BENEFITS)	Emergency response Fire weather Flash floods Safety of life	Emergency management Flood and storm warnings Irrigation scheduling Short-term planning	Emergency preparedness Water resource and crop management	Drought preparedness and management	Adaptation measures, Biodiversity and ecosystem conservation Resource development	Emissions reduction Strategic resource planning	
Drivers of variability and Change (SCIENCE)	Fronts, Convection, Local storms	Fronts Large scale weather systems Cut off lows Tropical cyclones	30-60 day tropical oscillations Monsoons Southern Annular Mode Blocking	Tropical air/sea interactions: EL Niño Indian Ocean Dipole	Anthropogenic forcing Other predictable sources not known for Southern Hemisphere – likely include deep ocean circulation and possibly sea ice	Anthropogenic forcing on mean climate	
Time scale →	Hours	Days	Weeks	Months	Years	Decades	Centuries



Australian Government
Bureau of Meteorology

Future Modelling

Provide a national approach to climate and weather prediction model development



Consistent with the strategy of *seamless prediction*



Future Systems

- Enhance our capacity for computing (100x or more by 2020)
 - Greater computing resources for ensemble model and assimilating data
 - Greater memory resources and performance
 - 3GB memory per Gflop of sustained computing
 - Greater storage resources and performance
 - Global parallel file systems (Lustre) scale-out in storage and bandwidth
 - Flash technology improves I/O bandwidth and IOPS
- Enhanced capability for climate and NWP computing
 - Whether the application can scale with growing number of cores?
 - Improvement in communications, and overlaps in computations
 - Hybrid models using MPI-OpenMP
 - Application I/O, parallel I/O or parallel data streams
 - Whether the application needs greater single processor capability?
 - Possibly new processors, new coding techniques and new languages
 - Or Wait for better tools



Computing Challenges

- **Issues with processor / memory imbalances**
 - Causes lower computational efficiencies (5% on Intel Nehalem)
 - Causes an increase in the number of cores for capability computing
 - Leads greater code scalability issues
 - Leads to increase power usage and costs
- **Increasing code inefficiencies**
 - MPI only codes in NWP and hi-res climate modelling is inefficient
 - Hybrid OpenMP/MPI is improving code scalability but not sufficient to greatly improve the (% of peak) achieved on a processor.
 - Question of code complexity with MPI only, and hybrid OpenMP/MPI increase that complexity.



Australian Government

Bureau of Meteorology

BUREAU OF METEOROLOGY

Thank you

Tim F. Pugh

Centre for Australian Weather and Climate Research

<http://www.cawcr.gov.au/>

<http://www.bom.gov.au/>

Phone: +61 3 9669 4345

Email: t.pugh@bom.gov.au



Climate Computing in higher resolution

Climate Computing - Intel Nehalem

# cores per node	8
Simulated years per Day	3.5
Watts per node	270
System Watts per compute node	405
Operating (\$ per kilowatt-hr)	\$0.20
Asset (\$ per cpu-hr)	\$0.13

Climate Computing - Intel Sandy Bridge

# cores per node	16
Simulated years per Day	3.5
Watts per node	270
System Watts per compute node	405
Operating (\$ per kilowatt-hr)	\$0.20
Asset (\$ per cpu-hr)	\$0.13

Component	Atmosphere	Ocean	Sea Ice	OASIS	System	Component	Atmosphere	Ocean	Sea Ice	OASIS	System
Configuration	N320L70	1/4 degree res	1/4 degree	MPI tasks		Configuration	N320L70	1/4 degree res	1/4 degree	MPI tasks	
# of Cores	1120	232	28	4	1384	# of Cores	640	128	28	4	800
Kilowatts per simulated year	259.2	53.7	6.5	0.9	320.3	Kilowatts per simulated year	111.2	22.2	4.9	0.7	139.0
\$Cost(power) per simulated year	\$51.84	\$10.74	\$1.30	\$0.19	\$64.06	\$Cost(power) per simulated year	\$22.24	\$4.45	\$0.97	\$0.14	\$27.81
\$Cost(asset) per simulated year	\$960.00	\$198.86	\$24.00	\$3.43	\$1,186.29	\$Cost(asset) per simulated year	\$548.57	\$109.71	\$24.00	\$3.43	\$685.71
Total Costs per simulated Year	\$1,011.84	\$209.60	\$25.30	\$3.61	\$1,250.35	Total Costs per simulated Year	\$570.82	\$114.16	\$24.97	\$3.57	\$713.52
% of Total	80.92%	16.76%	2.02%	0.29%		% of Total	80.00%	16.00%	3.50%	0.50%	

Comparison of high resolution climate computing on existing Intel Nehalem and estimates for Intel Sandy Bridge.

42% reduction in costs

Asset and power costs are assumed to be the same.



ACCESS – CMIP5 “Core” simulations for high resolution, Intel Sandy Bridge



Experiment	Length	Ens.
Preindustrial Control	500 yr (~200 yr spin up)	1
Historical	1850-2005	1
AMIP	1979 (or earlier?) - 2008	1
RCPs 4.5, 8.5	2006-2100	1
1%/yr CO ₂ to 4x	140	1
AGCM + control SSTs	30	1
AGCM + control SSTs + 4x CO ₂	30	1
4x CO ₂	150	1

AR5 total costs is \$806,990 and consume 157,240 kw-hrs
Actual costs, multiple by the number of unsuccessful runs



Challenges in Weather and Climate

Weather

- Challenge: 24-hour global coverage of atmosphere, land, ocean, sea ice, and wave observations.
- Challenge: Improving the model's scalability with decreasing spatial resolution and increasing MPI tasks/core counts.
- Challenge: Complete the assimilation analysis and model time integration within the forecast time window.
- Challenge: Managing the rapidly increasing data and storage volumes.
- Challenge: Managing the compute and storage infrastructure including human resources

Climate

- Challenge: Managing the complexity of interactions and verifying and validating the individual model components and system together.
- Challenge: Increasing the spatial resolution of the models to resolve important dynamical scales.
- Challenge: Improving the software's scalability with increasing spatial resolution and core counts.
- Challenge: Improving the software's efficiency and runtime consistency at high core counts. (system jitter/interference)
- Challenge: Managing the rapidly increasing data and storage volumes.
- Challenge: Managing the infrastructure and human resources.