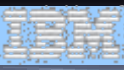




IBM Strategy for Weather & Climate

Jim Edwards – IBM Systems and Technology Group





Supercomputing for Climate and Weather Forecasting

FOCUSING ON THE PROBLEMS OF TODAY AND THE FUTURE



- **Challenges facing Climate & Weather**



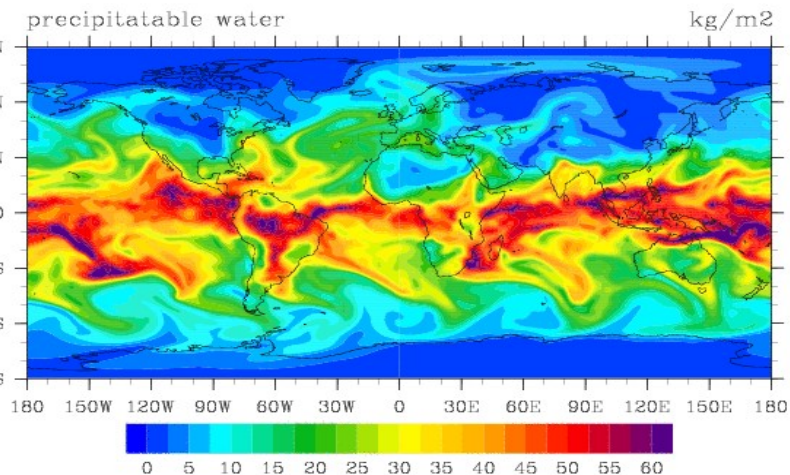
- **IBM Solutions**



- **IBM Experience**



- **POWER 7**





Supercomputing for Climate and Weather Forecasting

FOCUSING ON THE PROBLEMS OF TODAY AND THE FUTURE



➤ The Challenge facing Climate & Weather

- Increasing demands for more accurate science
-
- Recreating CCSM as a Petascale Climate Model
-
- HPC Power Consumption
-
-



Increasing demands for more accurate science

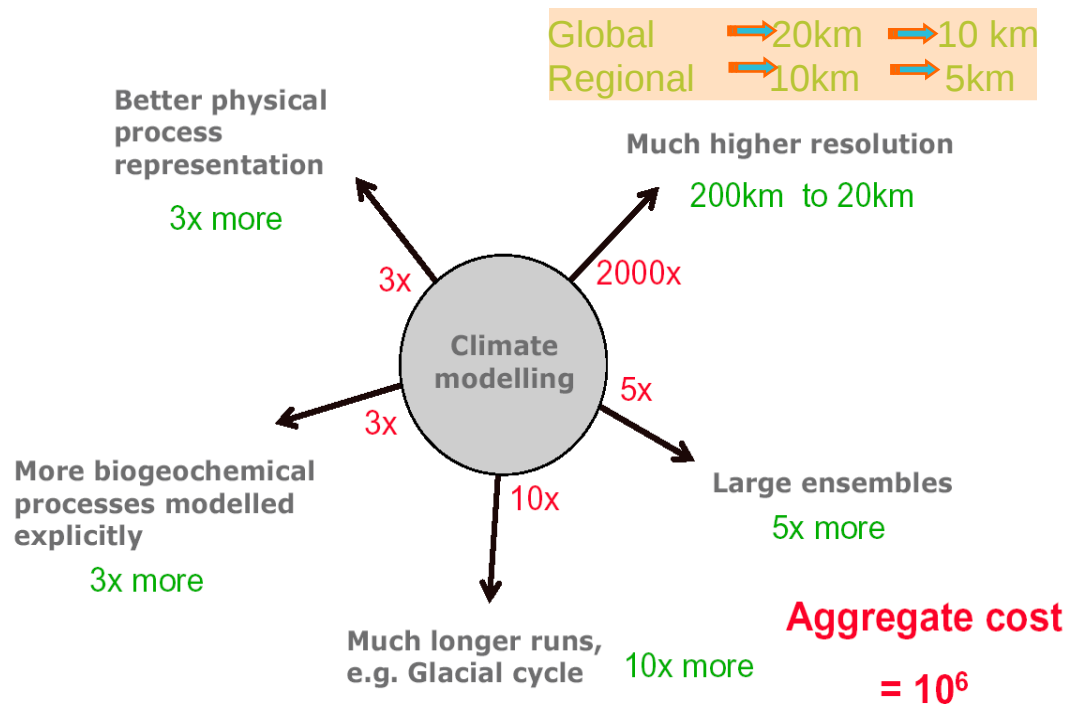
The Challenge facing Climate & Weather

Scientific demands are driving huge increases in capacity

Application scaling is forcing more ensemble solutions

Scaling is still a key factor in time sensitive production runs

Little sign of this slowing in next decade

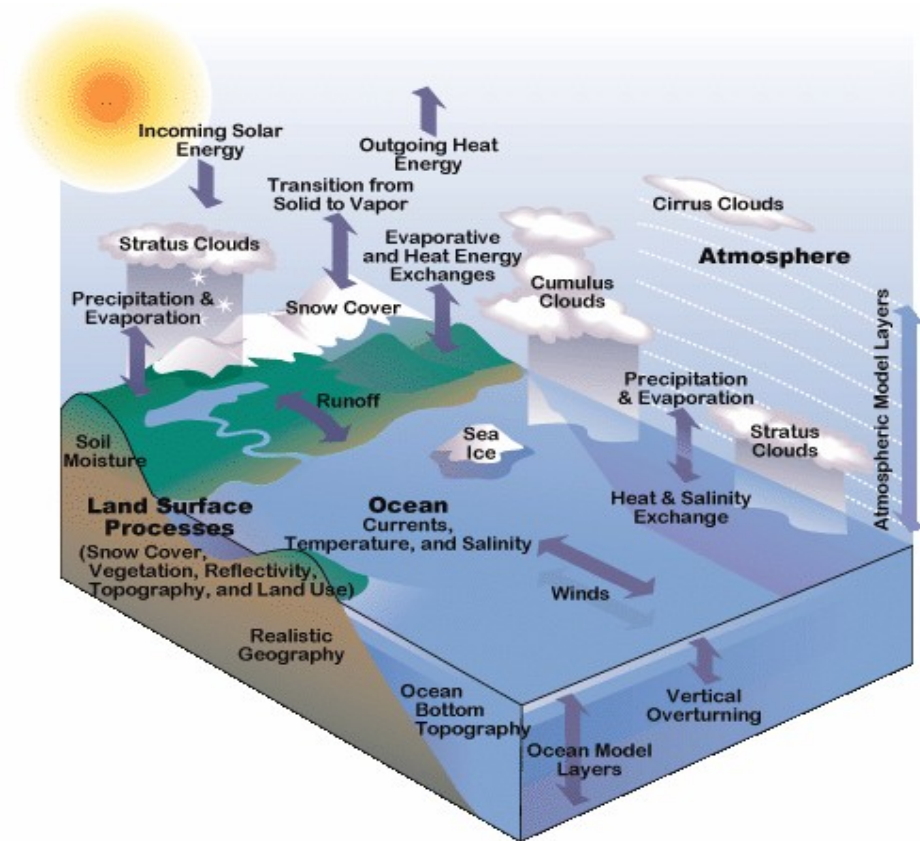


- **Typical Customer**
 - 7-10 X performance by 2015
 - 35-50 X performance by 2020



Achieving Petascale Performance in the NCAR Community Climate Systems Model

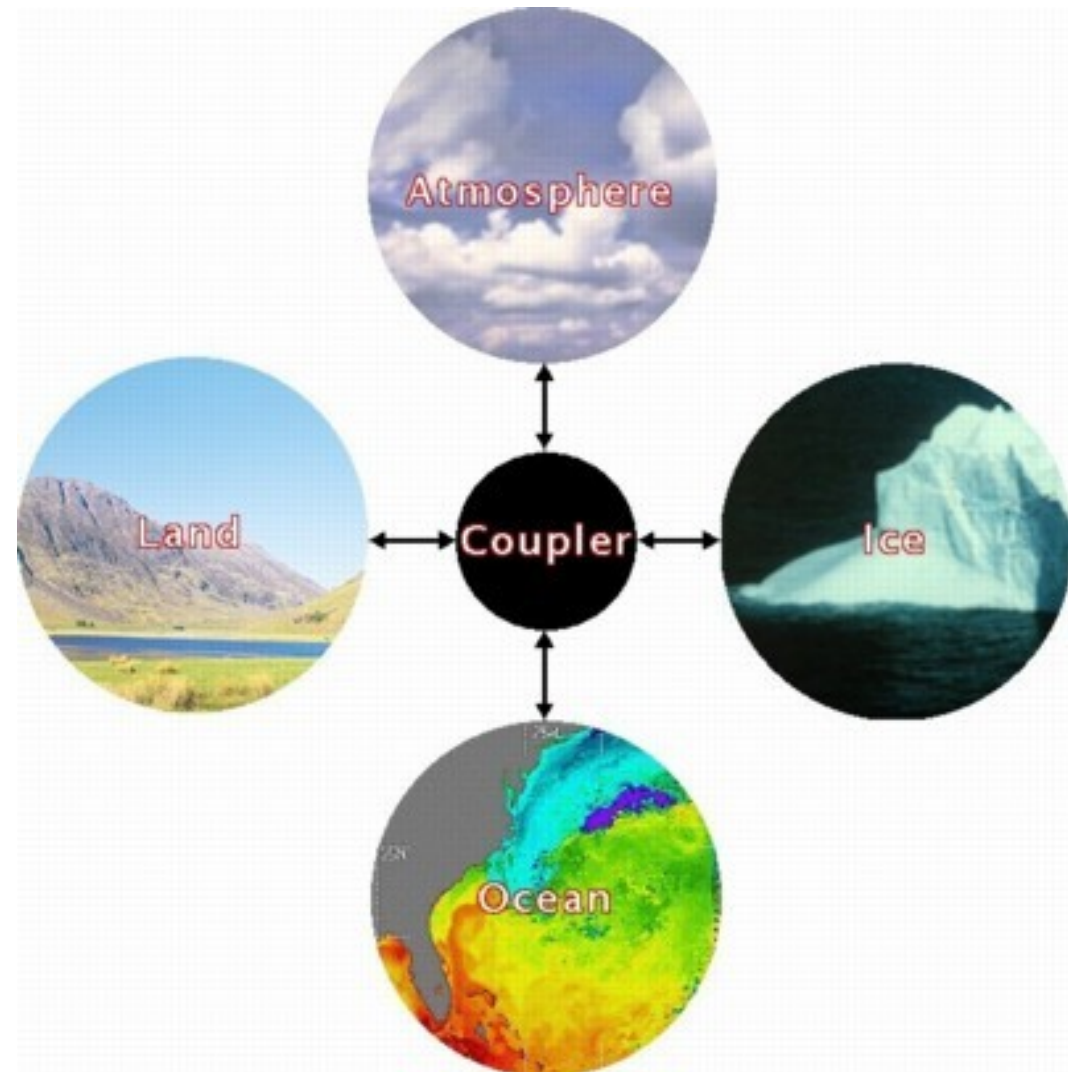
- CCSM
 - Fully coupled climate system with atmosphere, ocean, land and sea ice components
 - Major contributor to the 2007 IPCC report awarded the Nobel Peace Prize w/ Al Gore
 - Currently scales to 100's of processors
 - Scaling potential: 100,000+ processors





CCSM Model Impediments to Petascale

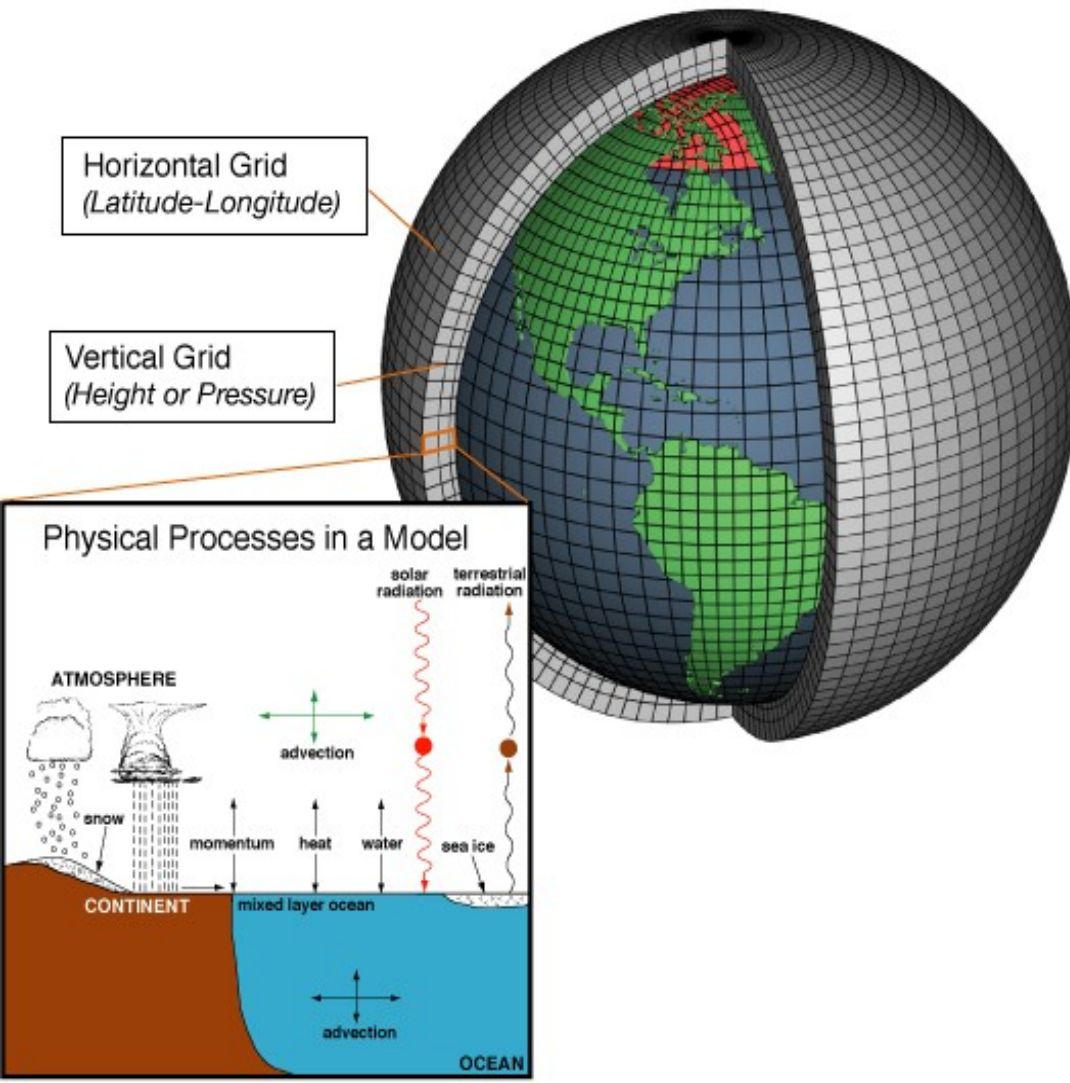
- Atmosphere
 - Dynamic Core implementation
- Ocean
 - Barotropic Solver
- All Components
 - I/O
 - Load Balancing



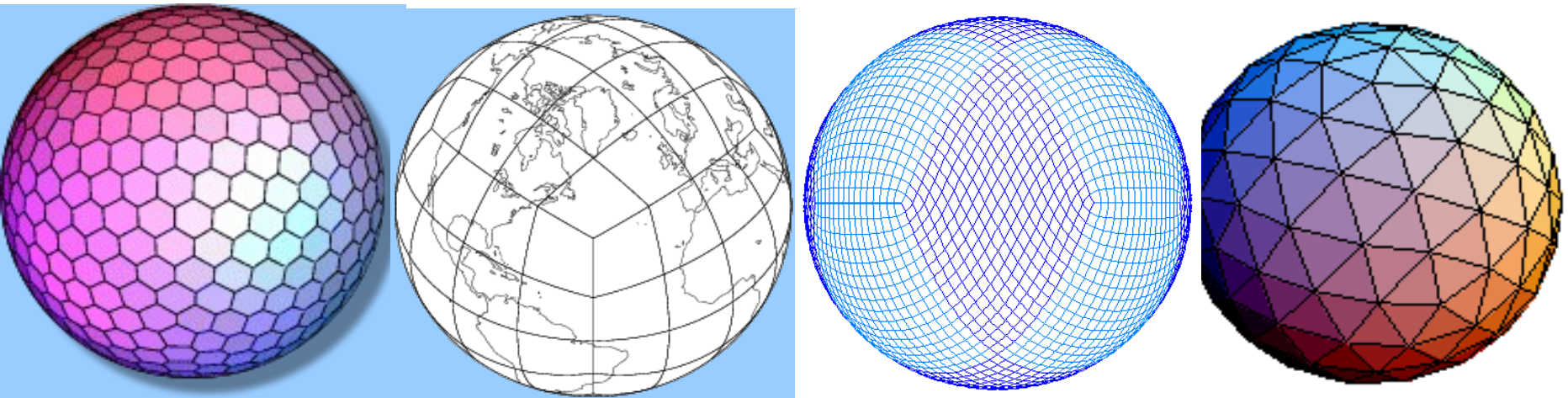


Atmospheric Dynamical Core Scalability

- Most atmospheric models use latitude – longitude grids
 - Well proven but have a pole problem
 - Pole problem has many good numerical solutions but these approaches tend to degrade parallel scalability



Atmospheric Dynamical Core Scalability

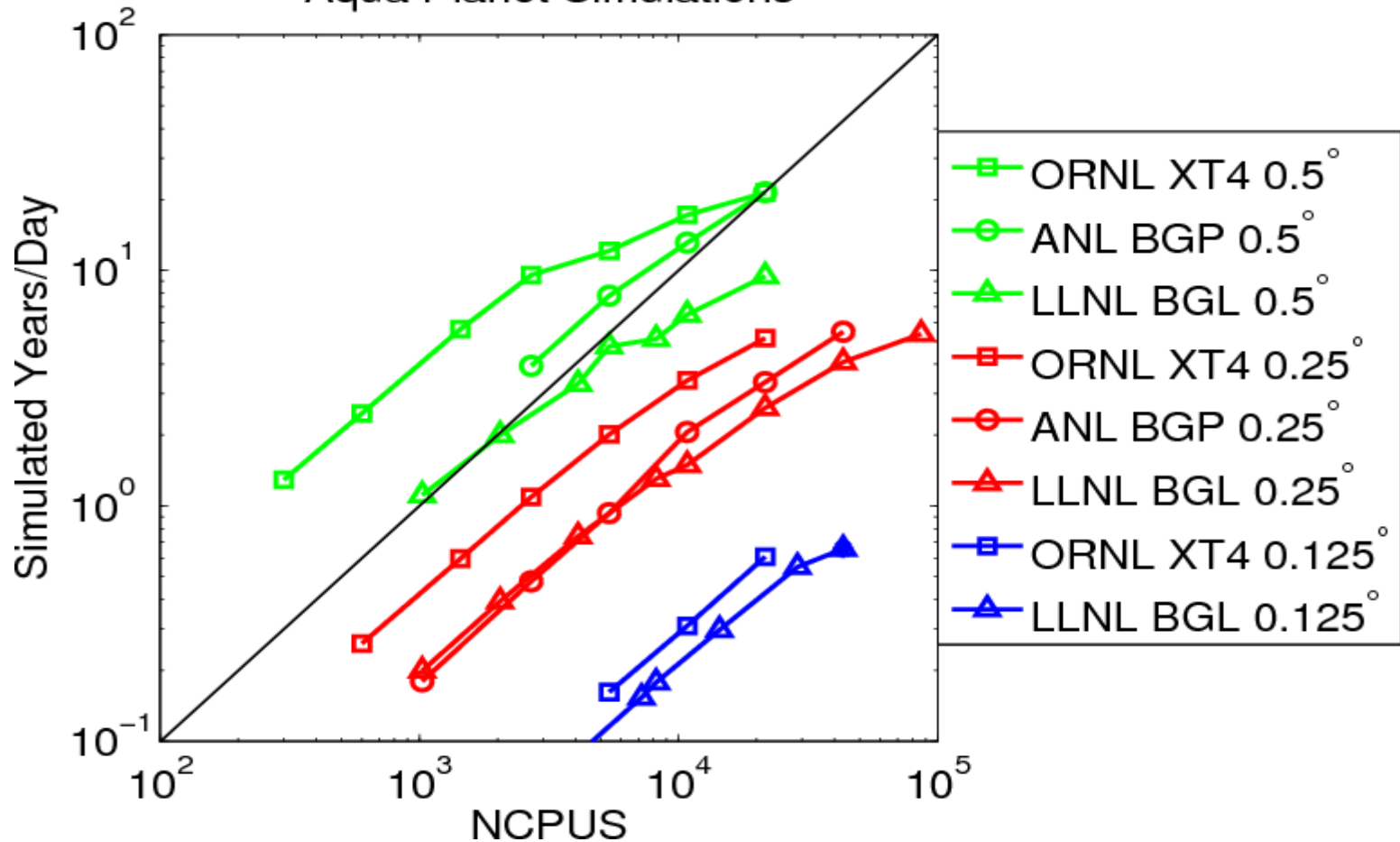


- Petascale dynamical cores:
 - Quasi-uniform grids avoid the pole problem
 - Can use full 2D domain decomposition in the horizontal
 - Equations can be solved explicitly with only nearest neighbor communication
 - BUT Numerical methods that perform as well as lat-lon methods remain a challenge
 - CCSM HOMME (Higher Order Mathematical Modeling Environment)
 - Conserve Mass, Energy and Vorticity in primitive variable formulation
 -



CCSM/HOMME Scalability

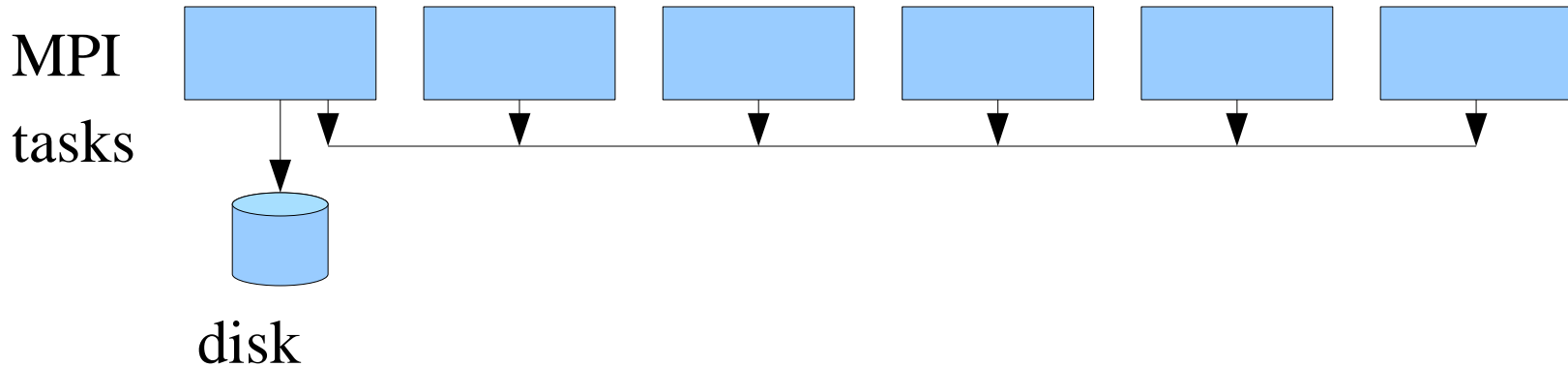
Aqua Planet Simulations



- Good scalability down to 1 horizontal element per processor
- On track to achieve petascale goal: 5 SYPD at 0.1 degree resolution



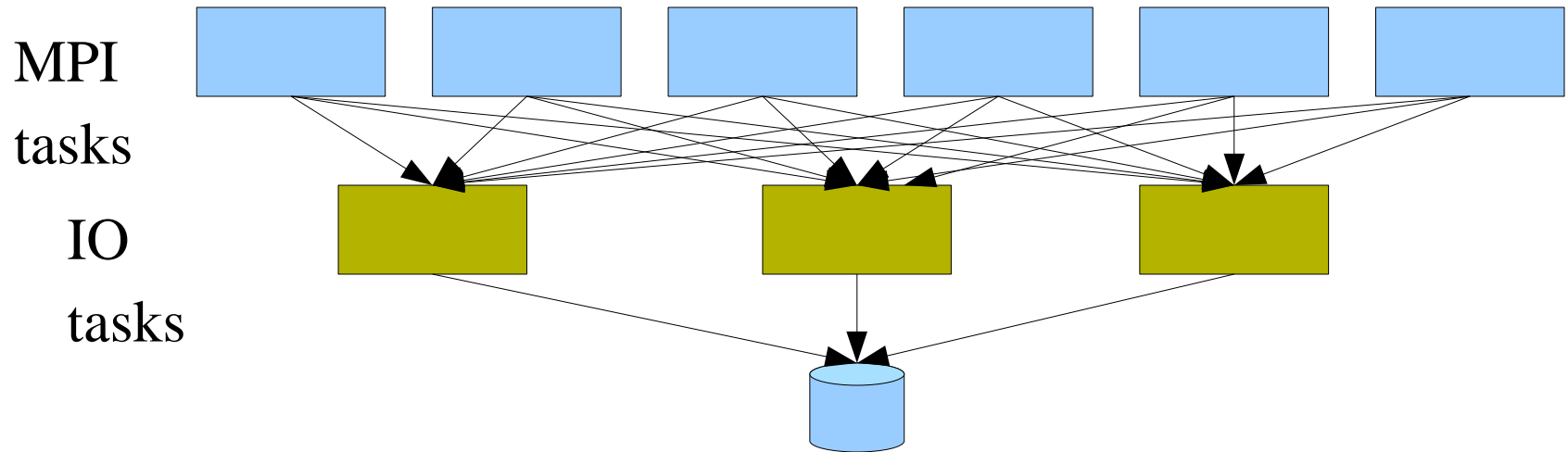
Improving CCSM I/O performance



- Current I/O method is serial
 - Performance is limited to single stream I/O rate
 - Potentially high memory requirement on “master” node
 - Uses netcdf format for a high degree of portability and compatibility with post processing tools



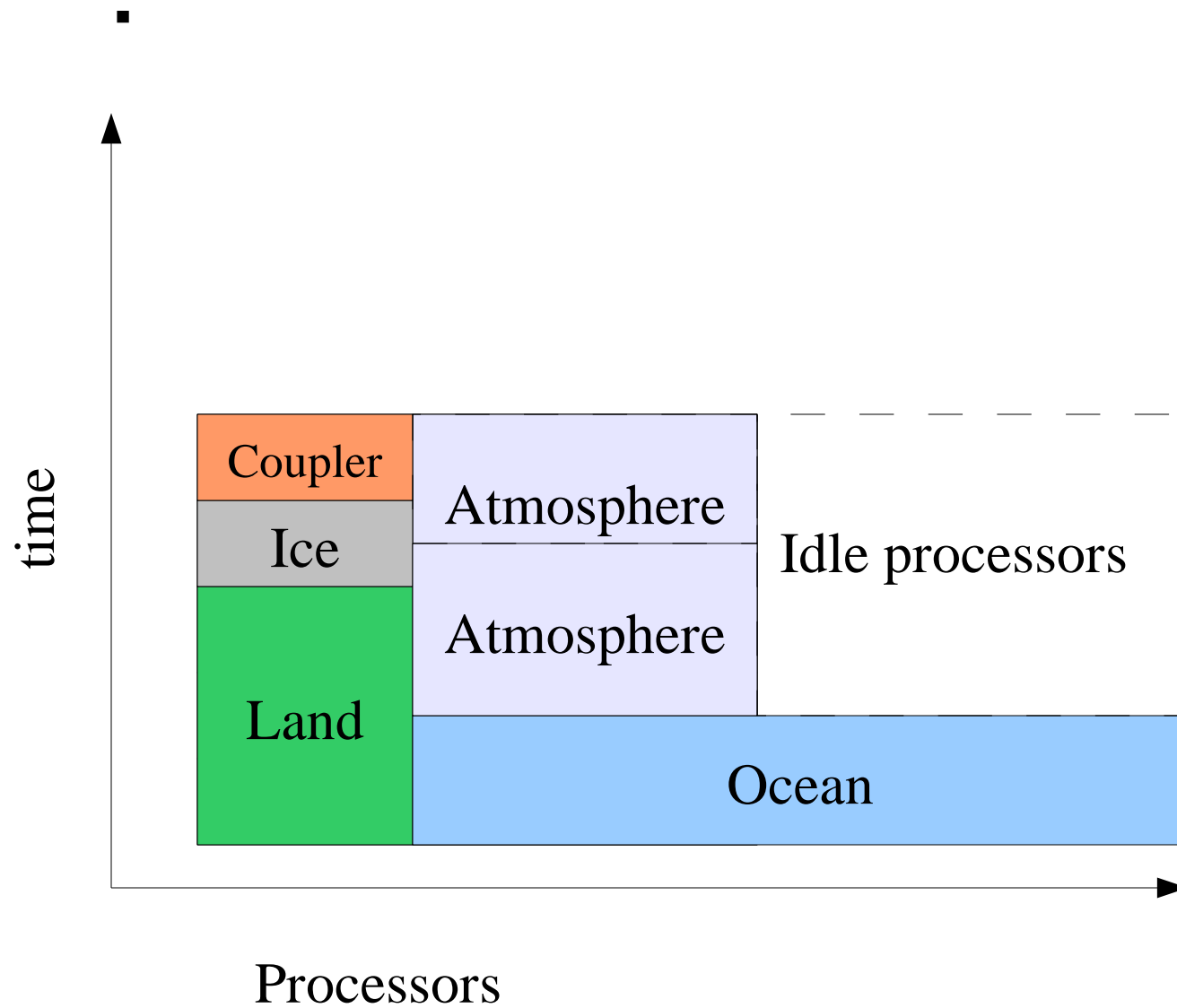
Improving CCSM I/O performance



- Parallel I/O subsystem to reorder data for optimal blocksize
- Can use netcdf or Pnetcdf (MPI-IO implementation of the netcdf format)
- 86,000 task BG/L runs would not otherwise have been possible
- <http://code.google.com/p/parallel-io>
-

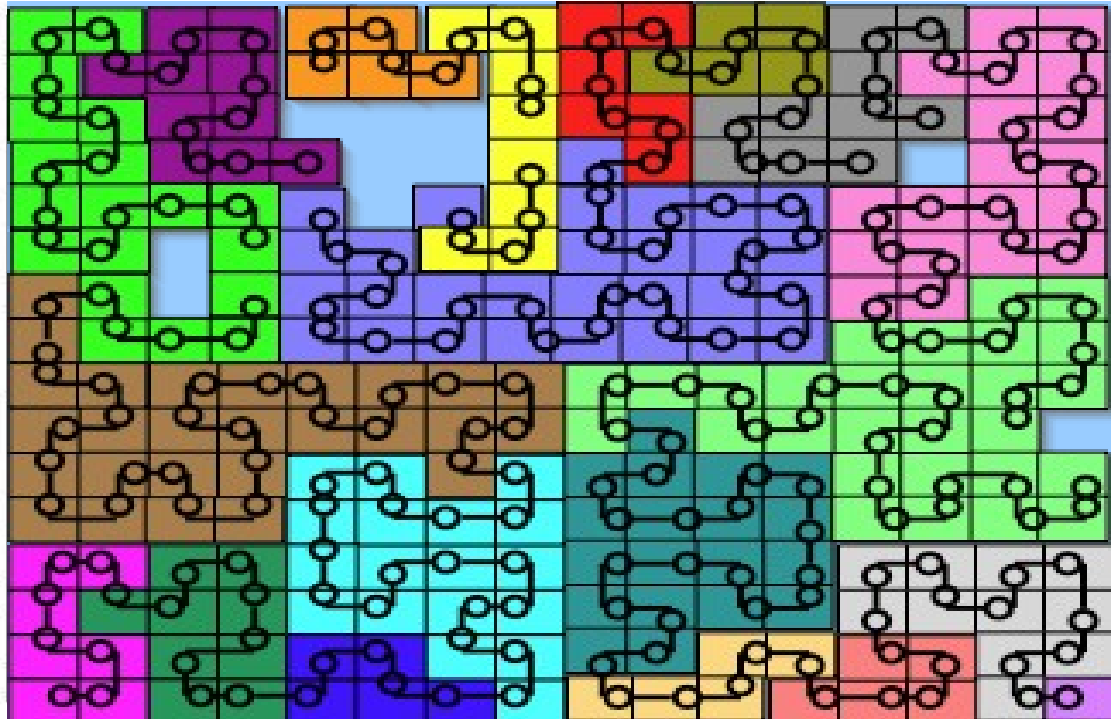


CCSM Load Balancing



CCSM Load Balancing

- Weighted space filling curves
 - Estimate work based on probability functions
 - Partition for equal amounts of work



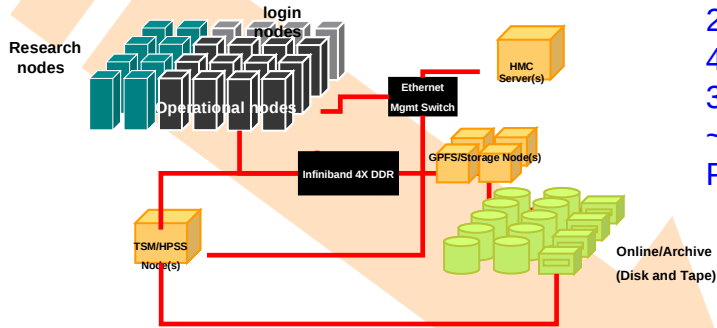


Power consumption & the Green Agenda

The Challenge facing Climate & Weather

- **As applications performance grows at >1.5x per year**
- **Energy use has to improve dramatically**
 - Space and cooling need to be managed
 - Application efficiency has to keep up with technology changes

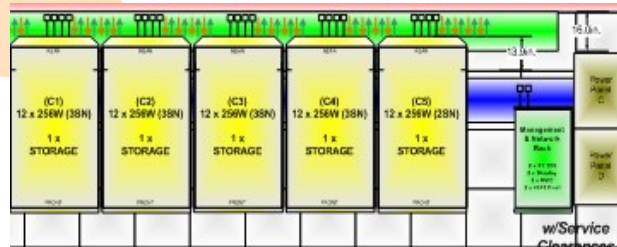
Typical P5 Weather cluster



22 racks
 4500 procs
 30 TFps (peak)
 ~650 KW
 Plus storage

Typical P7 Weather cluster

5-6 racks
 15,000 procs
 450 TFps (peak)
 ~850KW
 Storage included





NCAR Wyoming Supercomputer Center



- Maximum energy efficiency, LEED platinum certification and achievement of the smallest possible carbon footprint are all goals of the NWSC project
- 4.5MW to raised floor with a goal of 100% renewable energy
- Learn more: <http://www.cisl.ucar.edu/nwsc/>



Production level availability

The Challenge facing Climate & Weather

➤ Reliability

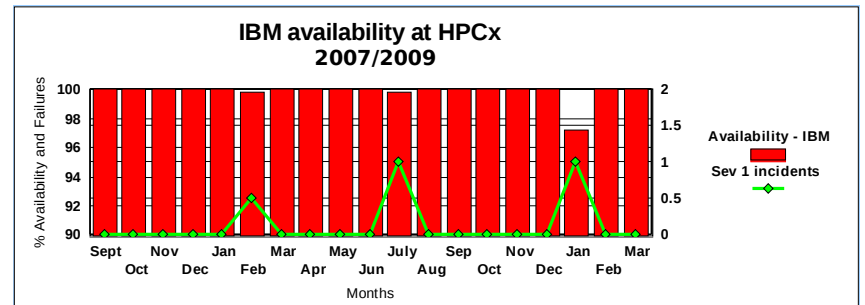
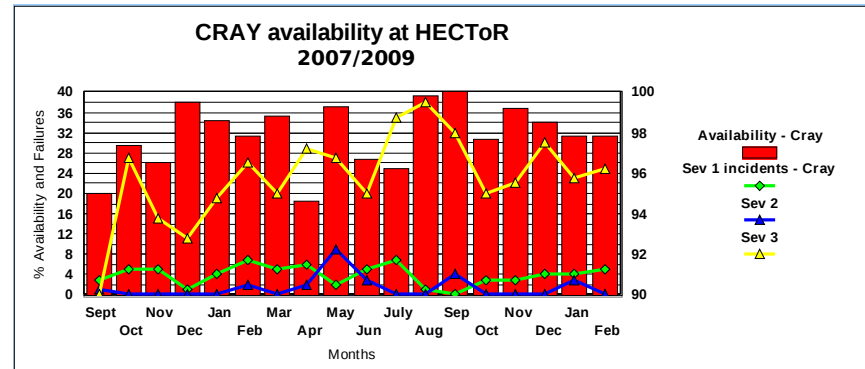
- Always available
- **Never miss a forecast**

➤ Performance

- Ability to run all current and future applications
- Fastest forecast production within budget
- Energy efficiency

➤ Hardware & Software capability

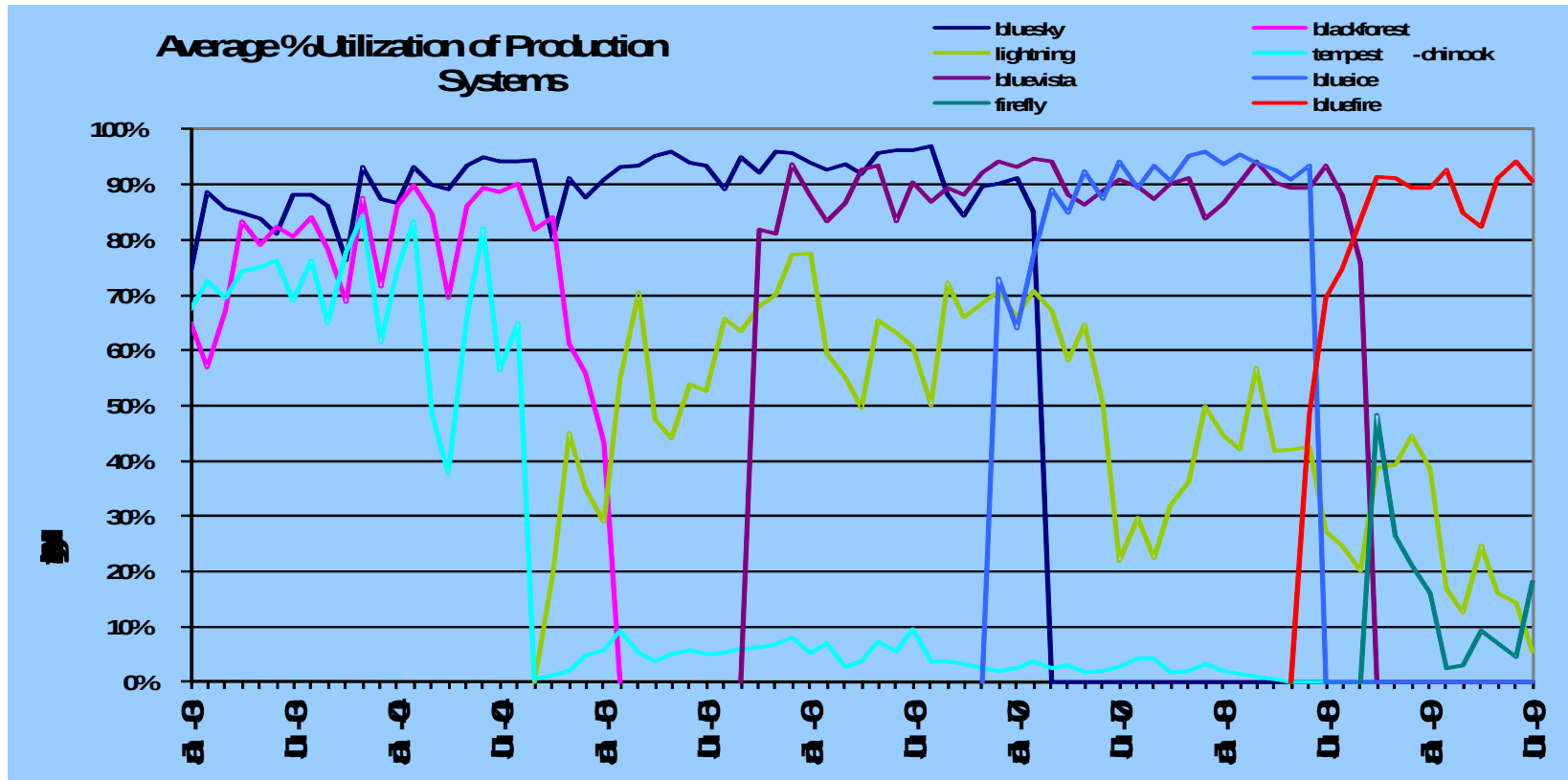
- Easy to use and manage
- Flexibility for users
- Strong support
- Future capability
- Strong user base



Real example from UK National Research Centre



NCAR systems Utilization



- Blueice (p5+) lifetime availability 99.2%
- Bluefire (p6) availability to date 97.4%



Power and energy aware job scheduling

•Power management in idling nodes

- Power down of idling nodes can affect overall productivity
- Leverage processor and OS features to put the idling node in minimum power consumption state

•Provide policy based power management

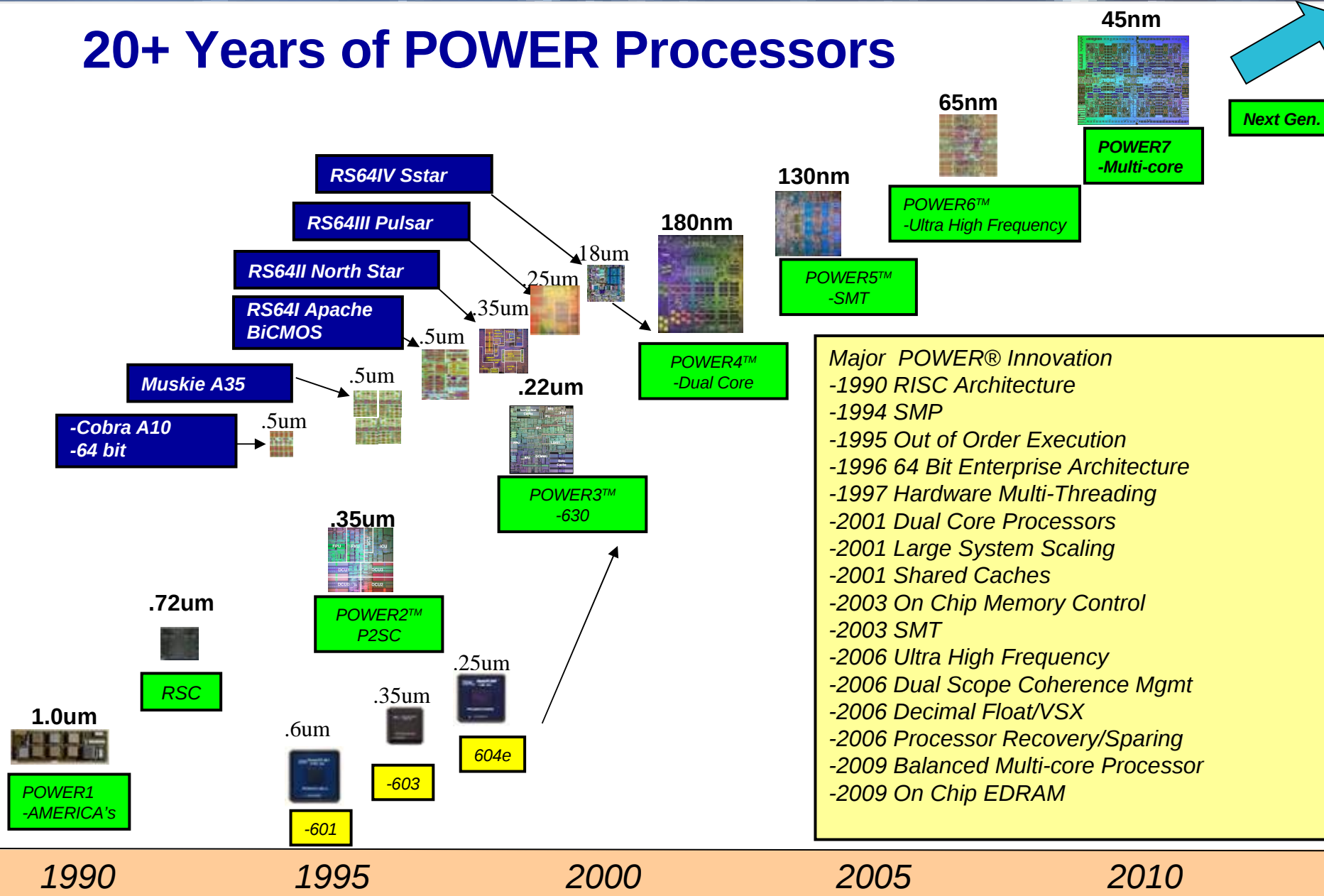
- Power management of either specific jobs or classes of jobs
- Policies can be set either by data center admins or users
 - e.g. lower priority jobs to be run in “power save” mode
 - e.g. jobs between 12 noon and 6PM to be run in “power save” mode
- Storing and reporting of power and energy consumption by job in DB
- Enable users and data center admins to define new policies based on previous usage

•Energy management

- Models to estimate power and energy consumption of jobs at different operating frequencies of the cores in a node
- Enables the job scheduler to decide what operating mode to use to minimize energy with least impact to performance



20+ Years of POWER Processors

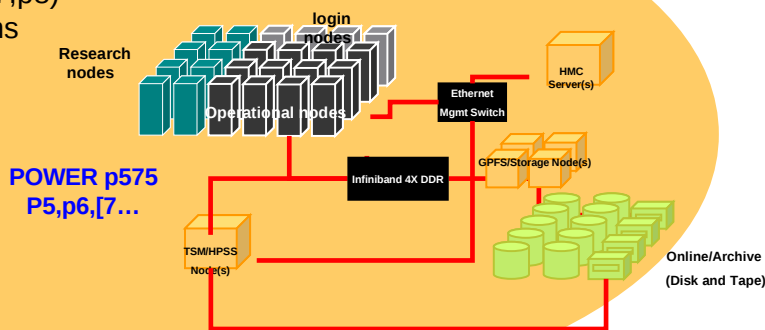




Typical solutions in Weather

IBM Solutions in Weather & Climate

- IBM Weather / Climate package
- POWER platform (p5,p6,p7,p8)
- Integrated software systems
- HPC support services



- Operational Systems
- Fast turnaround of forecast
- Ultra high availability
- Highly scalable
-
-
-
-
- Scientific Research
- Interoperability with production system
- Mix of Capability & Capacity
- Dynamic interoperability with
 - Operations

- Research Collaborations
- Application scaling
- Hybrid technologies
-
- iDataPlex x86
-
- Blue Gene
-
- Cell



BlueGene



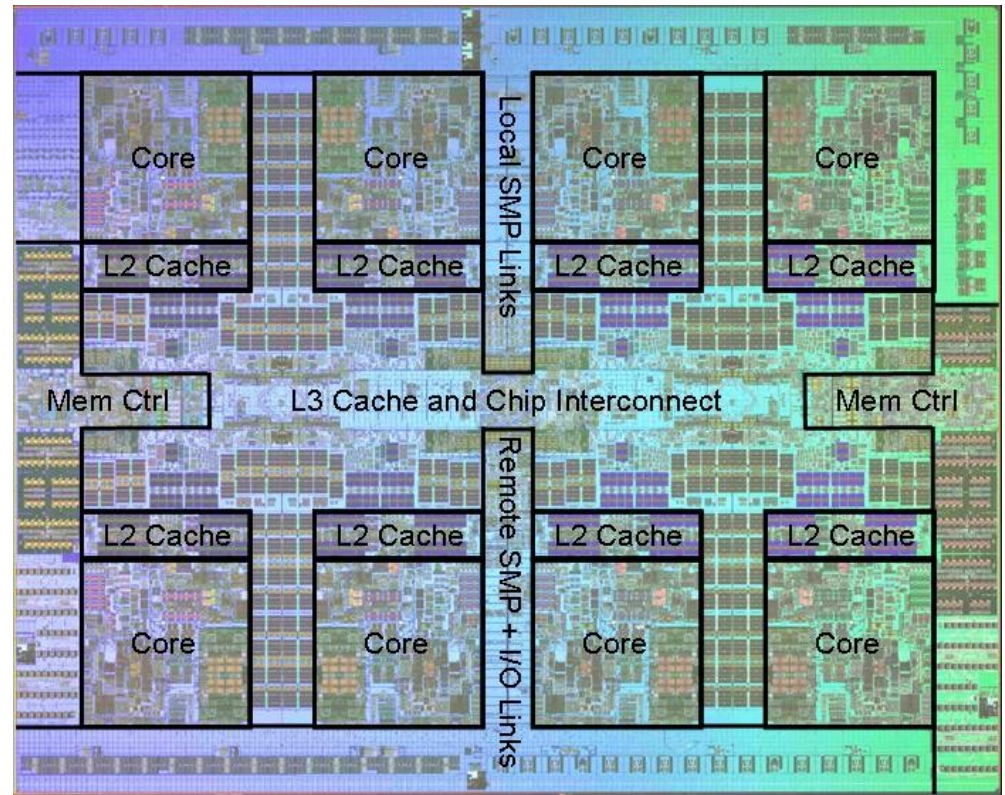
iDataPlex Blade

-
-
-
-
-
-
- Systems Research
- Modeling 2015 – 2020 systems
- 5 year + outlook
- Enabling ultra-scaling of applications
-



POWER7 Processor Chip

- 567mm² Technology: 45nm lithography, Cu, SOI, eDRAM
- 1.2B transistors
- Equivalent function of 2.7B
- eDRAM efficiency
- Eight processor cores
- 12 execution units per core
- 4 Way SMT per core
- 32 Threads per chip
- 256KB L2 per core
- 32MB on chip eDRAM shared L3
- Dual DDR3 Memory Controllers
- 100GB/s Memory bandwidth per chip sustained
- Scalability up to 32 Sockets
- 360GB/s SMP bandwidth/chip
- 20,000 coherent operations in flight
- Advanced pre-fetching Data and Instruction
- Binary Compatibility with POWER6
-



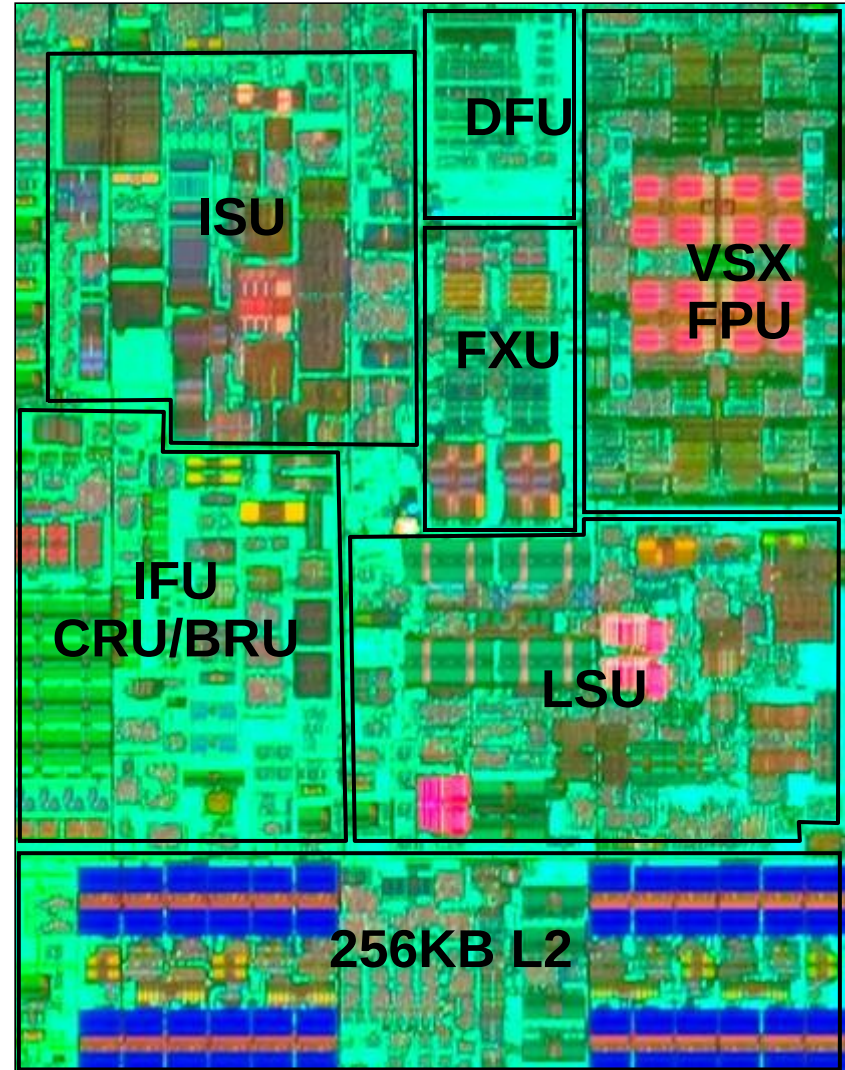
* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.



POWER7: Core

Execution Units

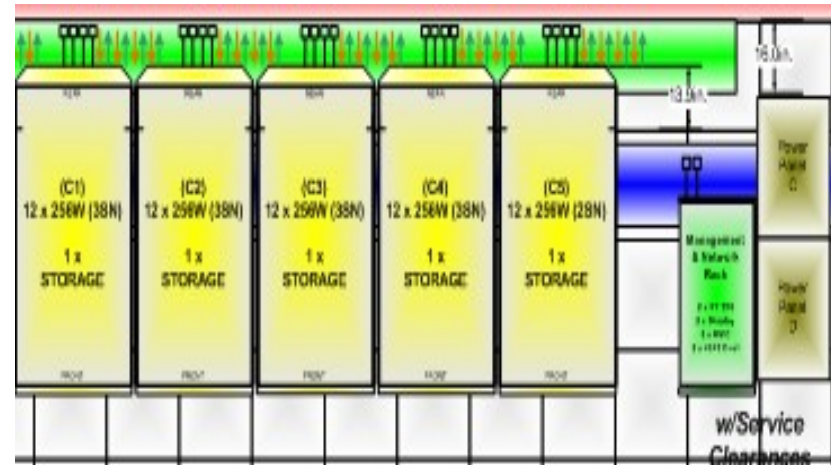
- 2 Fixed point units
- 2 Load store units
- 4 Double precision floating point
- 1 Branch
- 1 Condition register





P7 - 575

- High end follow on to p575
- Radically different implementation
- Using BlueGene like integrated technologies
- 32 way “Virtual” SMP node
- 256 core per drawer
- 3072 processors per rack
- Integrated switch



- Extra wide rack, very dense packaging
- 95% water cooled

•



Growing social impact of weather & climate predictions

L.A. Fire Sep 1, 2009

Taiwan typhoon Aug 11, 2009



Worldwide HPC Weather & Environmental Sites

