Running Operational Weather Models on FNMOC's Linux Cluster

Roger A. Stocker
R. Michael Clancy

Overview of Operations

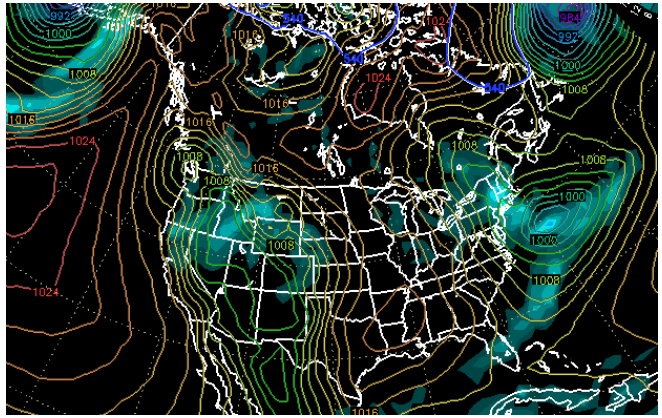
Operations

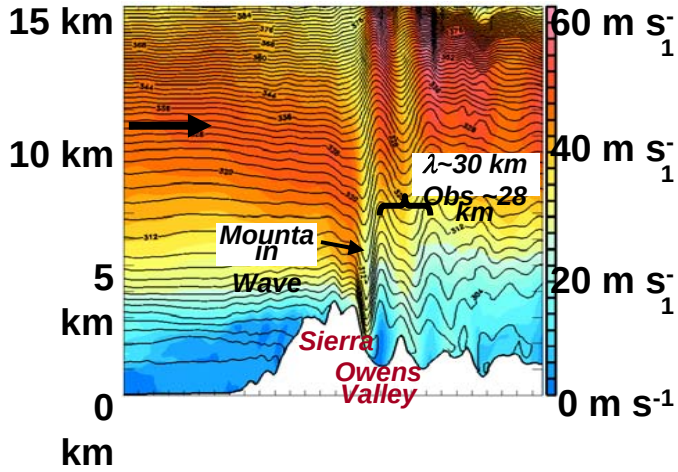
- **Operations Center**
  - Manned 24x7 by a team of military and civilian watch standers
  - Focused on operational mission support, response to requests for special support and products, and customer liaison for DoD operations worldwide
  - Joint Task Force capable
  - The Navy's Worldwide Meteorology/Oceanography Operations Watch
  - Operates at UNCLAS, CONFIDENTIAL and SECRET levels
  - SUBWEAX mission at SECRET level
- **Sensitive Compartmented Information Facility (SCIF)**
  - Extension of Ops Center
  - Operational communications (including secure video teleconferencing), tasking and processing elevated to the TS/SCI level if needed
  - Includes significant supercomputer capacity at TS/SCI level
  - SUBWEAX mission at TS/SCI level
- **Ops Run**
  - Scheduled and on-demand 24x7 production
  - 6 million meteorological observations and 2 million products each day
  - 15 million lines of code and ~16,000 job executions per day
  - Highly automated and VERY reliable

Models

- **Fleet Numerical operates a highly integrated and cohesive suite of global, regional and local state-of-the-art weather and ocean models:**
  - **Navy Operational Global Atmospheric Prediction System (NOGAPS)**
  - **Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS)**
  - **Navy Atmospheric Variational Data Assimilation System (NAVDAS)**
  - **Navy Aerosol Analysis and Prediction System (NAAPS)**
  - **GFDN Tropical Cyclone Model**
  - **WaveWatch 3 (WW3) Ocean Wave Model**
  - **Navy Coupled Ocean Data Assimilation (NCODA) System**
  - **Ensemble Forecast System (EFS)**

Surface Pressure and Clouds Predicted by NOGAPS

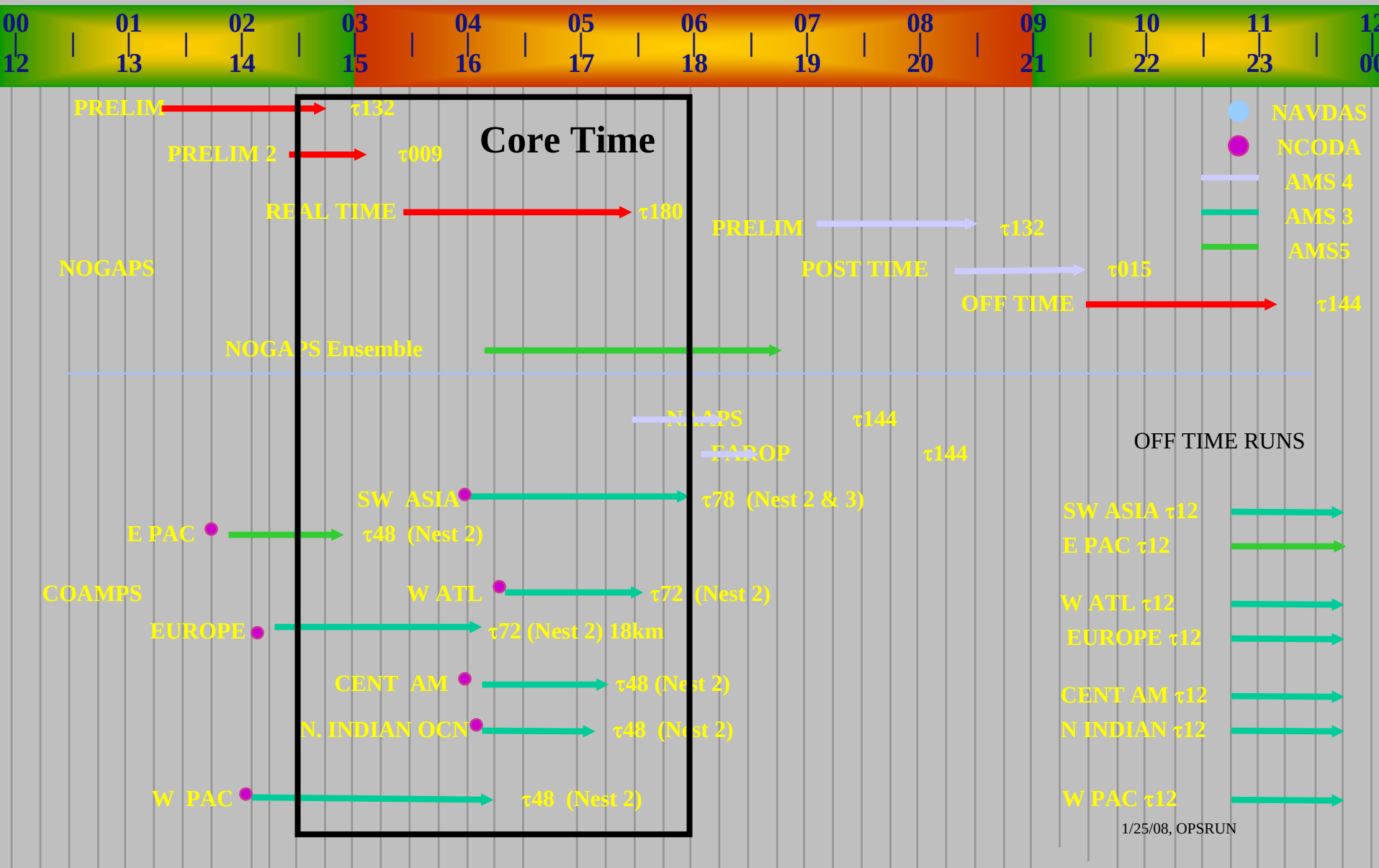




Cross section of temperature and wind speeds from COAMPS showing mountain waves over the Sierras

Current Operational Run

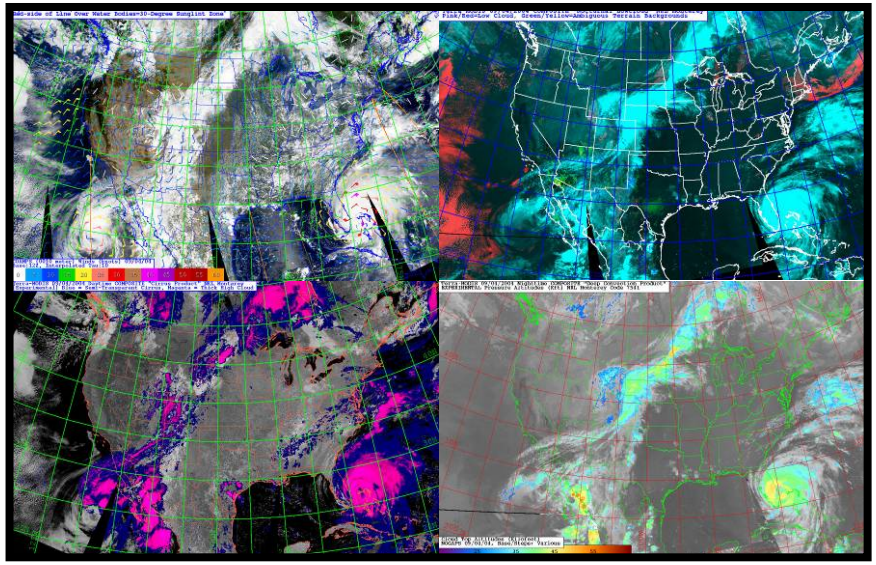
ATMOSPHERIC MODELS



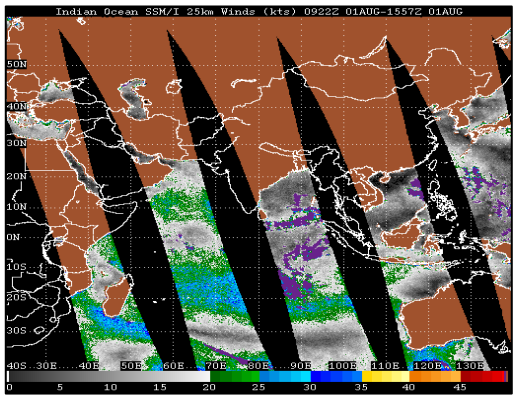
1/25/08, OPSRUN

Satellite Products

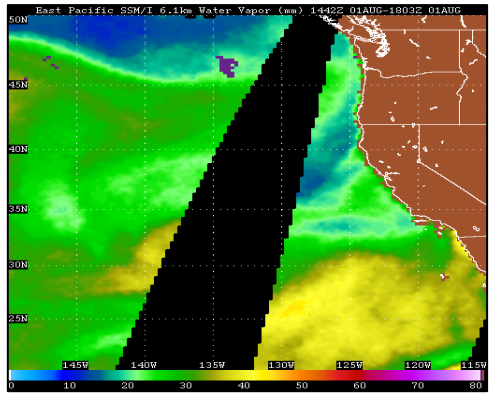
- SATFOCUS
- SSMI and SSMI/S
- Scatterometer
- Tropical Cyclone Web Page
- Target Area METOC (TAM)
- Tactically Enhanced Satellite Imagery (TESI)



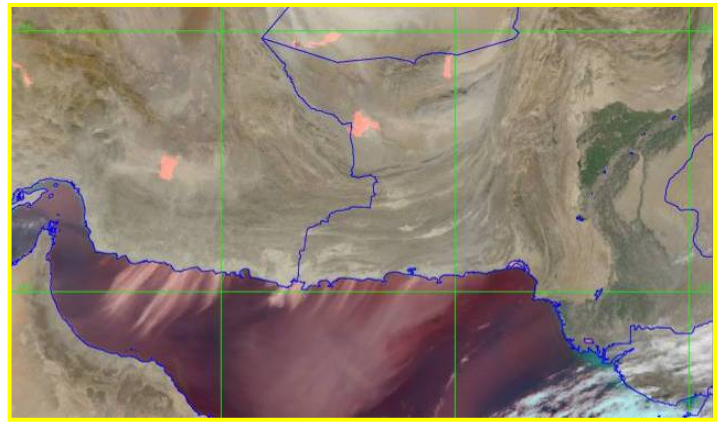
Example SATFOCUS Products



SSM/I Wind Speed



SSM/I Water Vapor



SATFOCUS Dust Enhancement Product

Overview of POPS

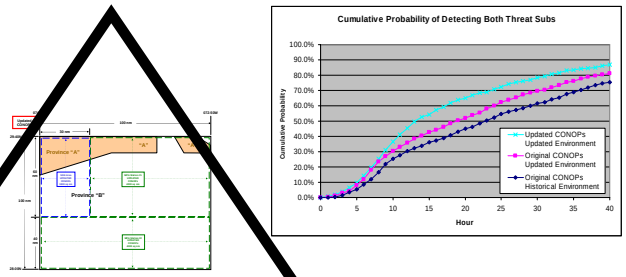
Primary Oceanographic Prediction System (POPS)

- **Provides HPC platforms to run models and applications at the UNCLAS, SECRET and TS/SCI Levels**
- **Has been traditionally composed of two Subsystems:**
  - **Analysis and Modeling Subsystem (AMS)**
    - **Models (NOGAPS, COAMPS, WW3, etc.)**
    - **Data Assimilation (NAVDAS, NAVDAS-AR)**
  - **Applications Transactions and Observations Subsystem (ATOS)**
    - **Model pre- and post-processing (e.g., observational data prep/QC, model output formatting, model visualization and verification, etc.)**
    - **Full range of satellite processing (e.g., SATFOCUS products, TCWeb Page, DMSP, TAM, etc.)**
    - **Full range of METOC applications and services (e.g., OPARS, WebSAR, CAGIPS, AREPS, TAWS, APS, ATCF, AOTSR, etc.)**
    - **Hosting of NOP Portal (single Web Presence for Naval Oceanography)**

Battlespace on Demand (BonD) and Current POPS Subsystems

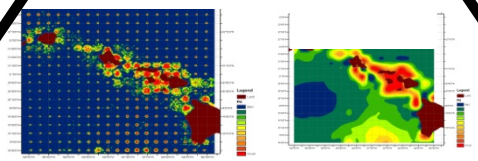
Forecast Battlespace

Tier 3 – the Decision Layer

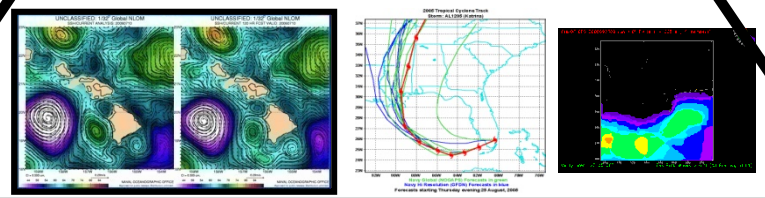


ATOS

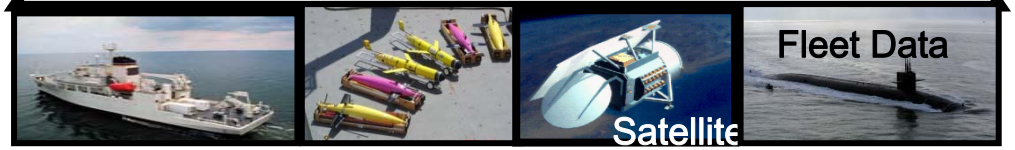
Tier 2 – the Performance Layer



Tier 1 – the Environment Layer



AMS



ATOS

Observational Data

POPS Architecture Strategy Going Forward

- **Drivers:**
  - Knowledge Centric CONOPS, requiring highly efficient reach-back operations, including low-latency on-demand modeling
  - Growing HPC dominance of Linux clusters based on commodity processors and commodity interconnects
  - Resource ($) Efficiencies
- **Solution:**
  - Combine AMS and ATOS functionality into a single system
  - Make the system 'vendor agnostic'
  - Use Linux-cluster technology based on commodity hardware
- **Advantages:**
  - Efficiency for reach-back operations and on-demand modeling
    - Shared file system and shared databases
    - Lower latency for on-demand model response
  - Capability to surge capacity back-and-forth among the AMS (BonD Tier 1) and ATOS (BonD Tiers 0, 2, 3) applications as needed
  - Cost effectiveness of Linux and commodity hardware vice proprietary operating systems and hardware
  - Cost savings by converging to a single operating system [Linux]
- **We call the combined AMS and ATOS functionality A2 (AMS + ATOS = A2)**

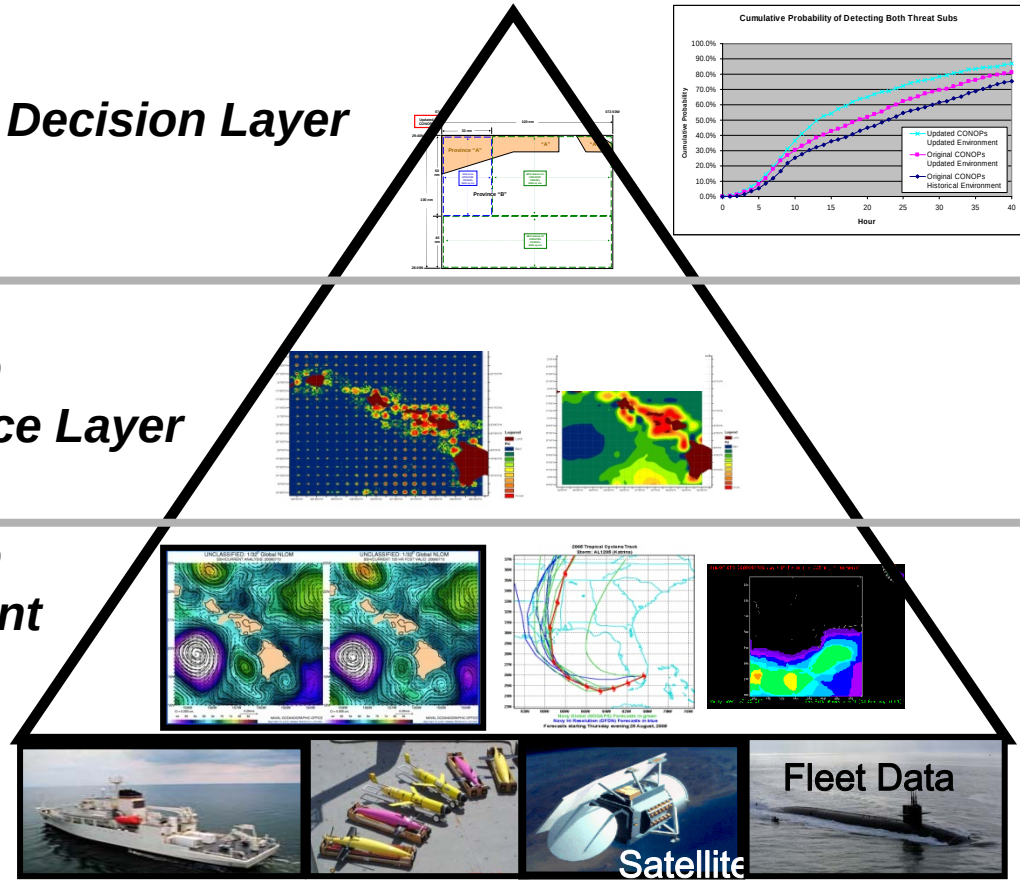
BonD and Future POPS

Forecast Battlespace

Tier 3 – the Decision Layer

Tier 2 – the Performance Layer

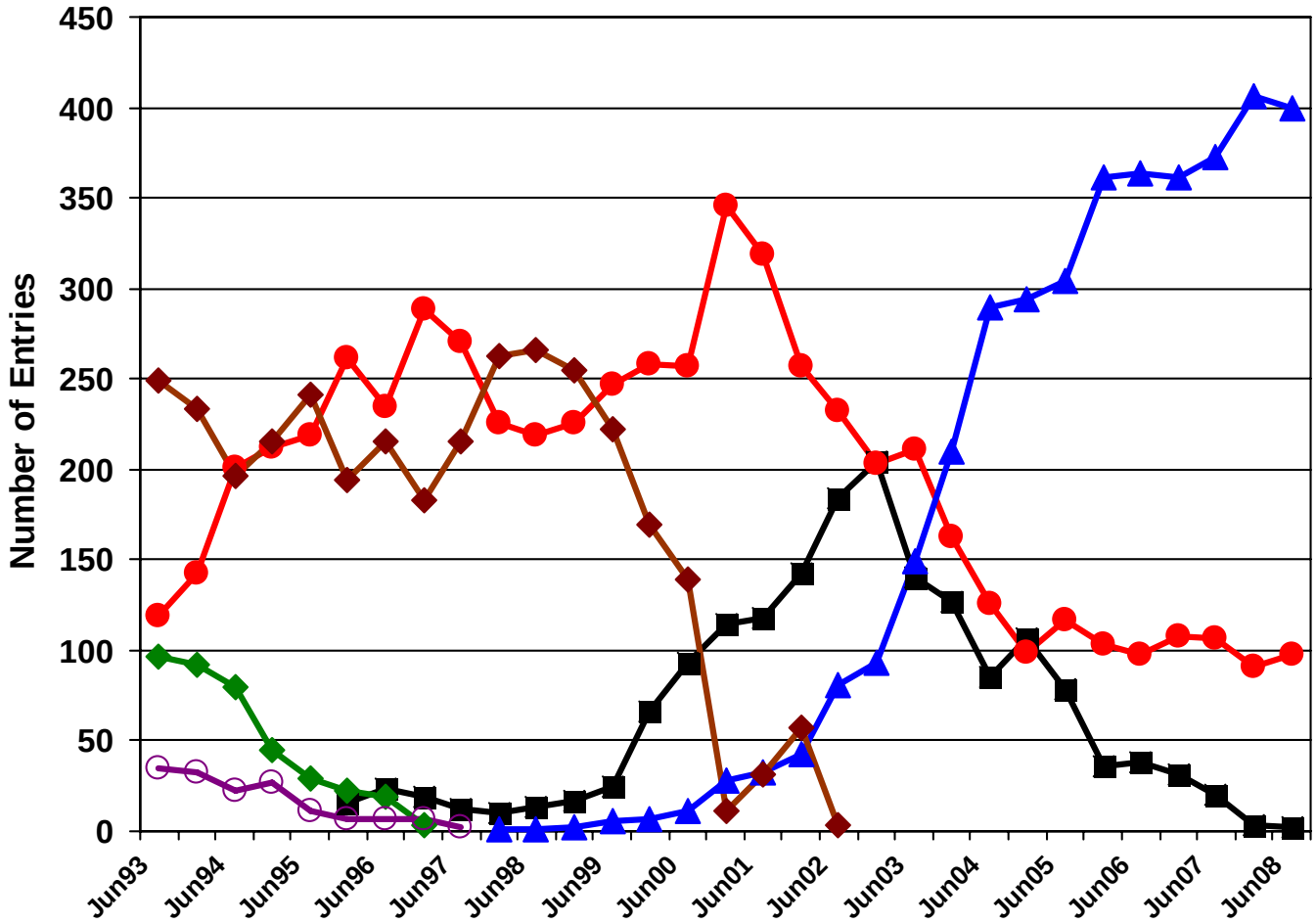
Tier 1 – the Environment Layer



A2

Observational Data

Trends in Architecture for HPC Systems on the TOP500 List



Definitions:

Constellation – Cluster of SMP nodes

MPP – Massively Parallel Processor

Cluster – Commodity processor and interconnect

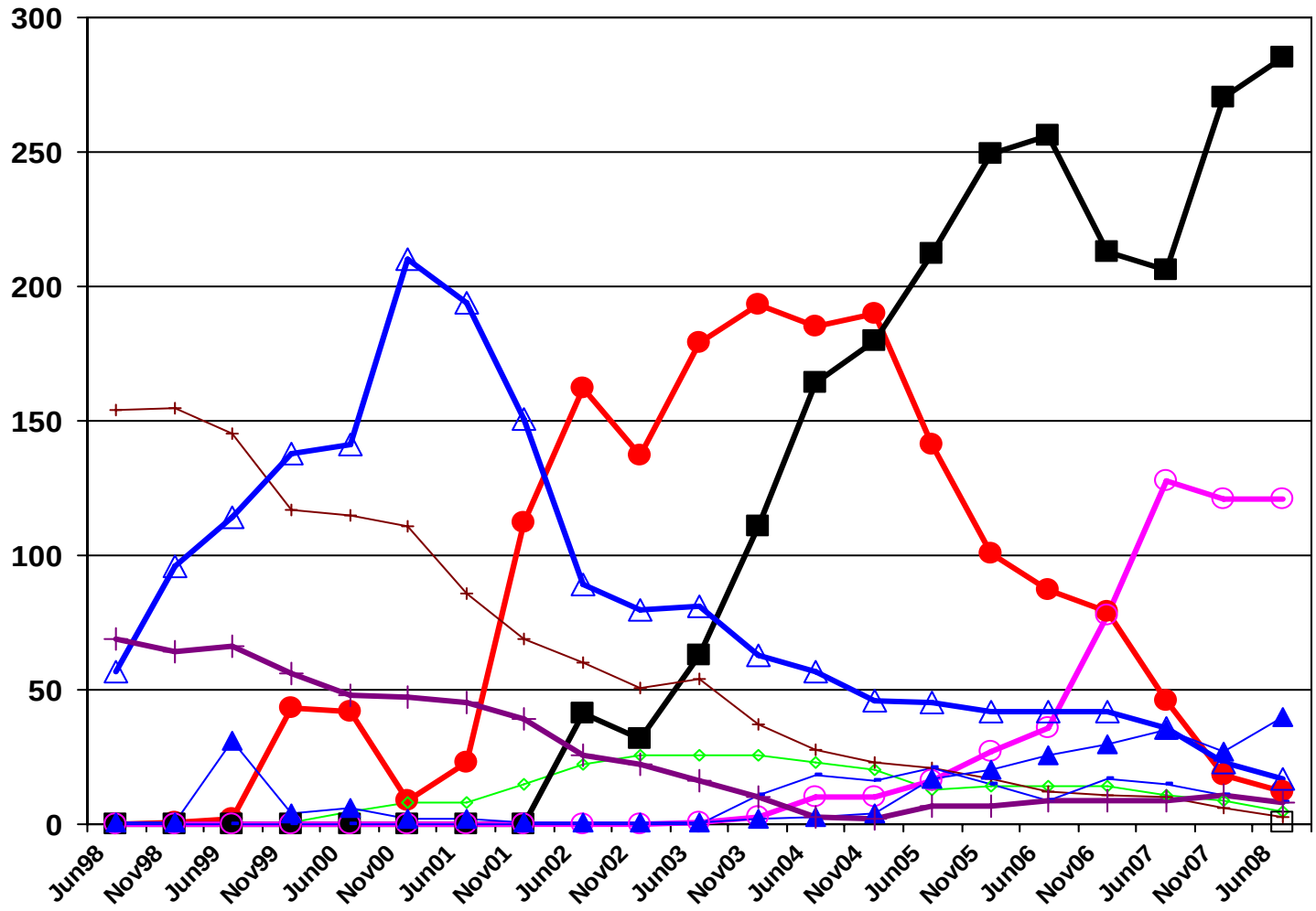
SMP – Symmetric Multi-Processor

Single Processor

SIMD – Single Instruction stream Multiple Data stream

Source: www.top500.org

Trends in Interconnect Technology for HPC Systems on the TOP500 List

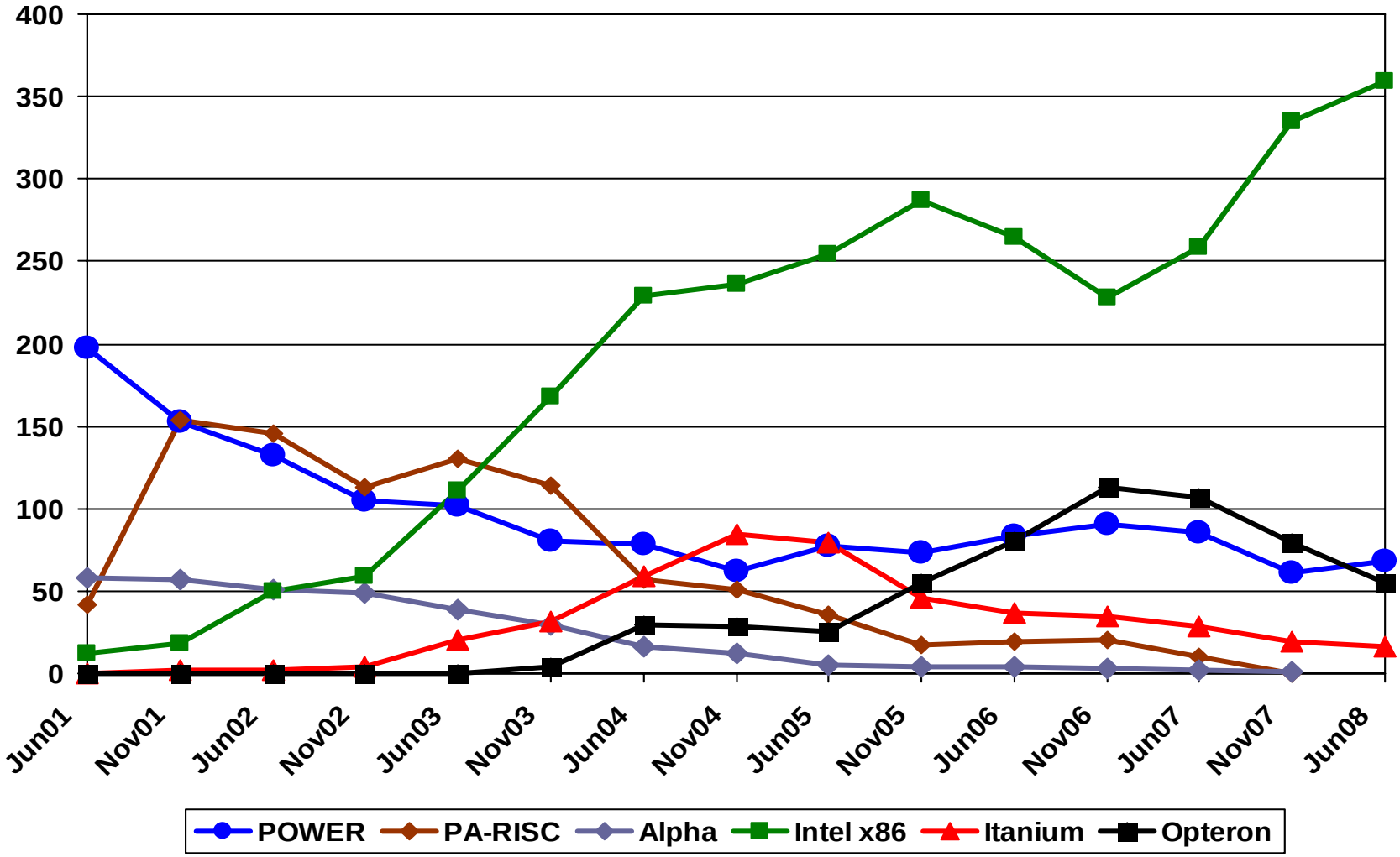


Notes:

1. only major interconnect families shown
2. Proprietary includes Blue Gene fabric

Source: www.top500.org

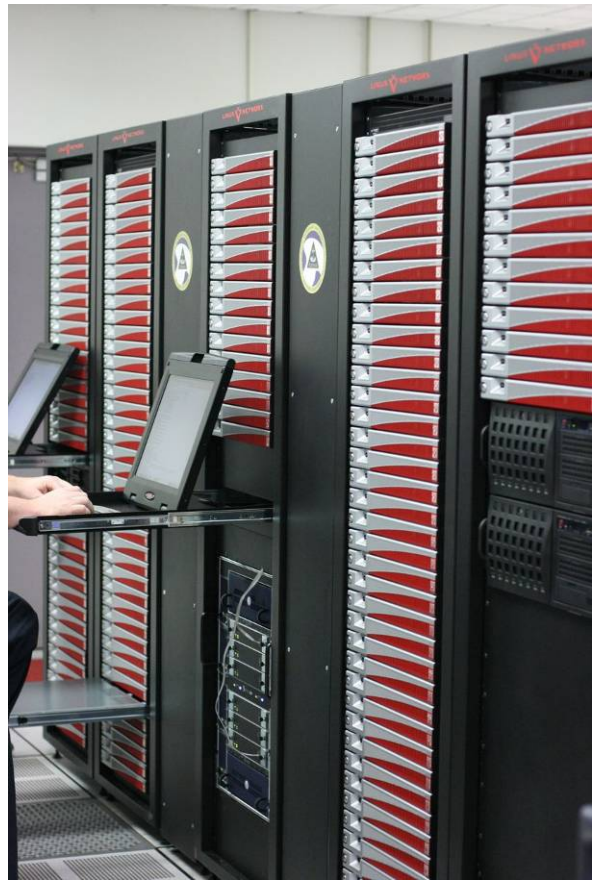
Trends in Processor Technology for HPC Systems on the TOP500 List



Source: www.top500.org

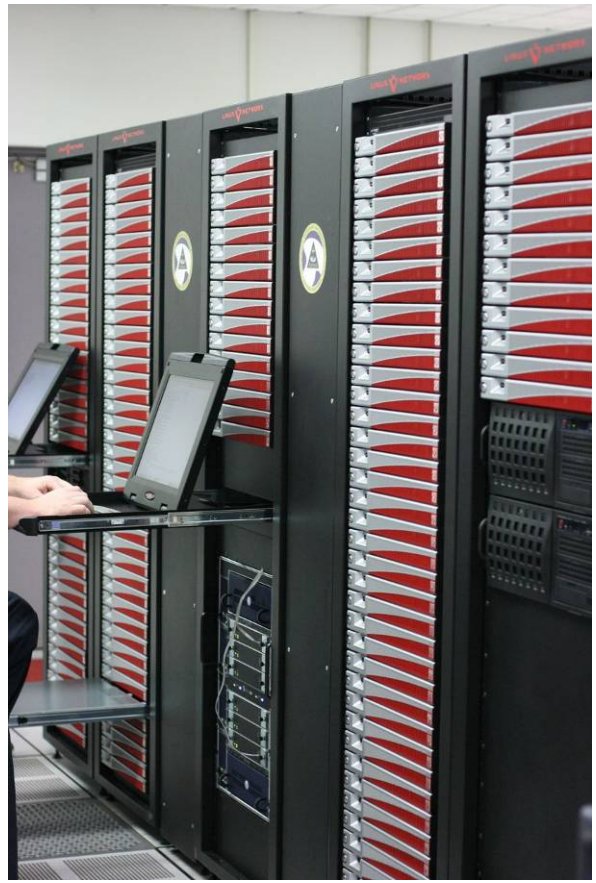
A2 Hardware Specifications

- **A2-0 (Opal) - Linux Cluster**
  - 232 nodes with 2 Intel Xeon 5100-Series “Woodcrest” dual-core processors per node (928 core processors total)
  - 1.86 TB memory (8 GB per node)
  - 10 TFLOPS peak speed
  - 35 TB disk space with 2 GB per second throughput
  - Double Data Rate Infiniband interconnect at 20 Gb per second
  - 40 Gb per second connection to external Cisco network
  - 4 Clearspeed floating point accelerator cards
- **A2-1 (Ruby) - Linux Cluster**
  - 146 nodes with 2 Intel Xeon 5300-Series “Clovertown” quad-core processors per node (1168 core processors total)
  - 1.17 TB memory (8 GB per node)
  - 12 TFLOPS peak speed
  - 40 TB disk space with 2 GB per second throughput
  - Double Data Rate Infiniband interconnect at 20 Gb per second
  - 40 Gb per second connection to external Cisco network
  - 2 GRU floating point accelerator cards



A2 Hardware Specifications

- **A2-2 (Topaz) - Linux Cluster**
  - **35 nodes with 2 Intel Xeon 5400-Series “Harpertown” quad-core processors per node (256 core processors total)**
  - **1.2+ TB Memory**
  - **3+ TFLOPS peak speed**
  - **16 TB disk space with 2 GB per second throughput**
  - **Double Data Rate Infiniband interconnect at 20 Gb per second**
  - **3 VMWare Nodes with 32GB/node**



A2 Middleware

- **Beyond the hardware components, A2 depends on a complicated infrastructure of supporting systems software (i.e., "middleware"):**
  - **Suse Enterprise Server Linux Operating System**
  - **GPFS: the Global Parallel File System, commercial software that provides a file system across all of the nodes of the system**
  - **ISIS: Fleet Numerical's in-house database designed to rapidly ingest and assimilate NWP model output files and external observations**
  - **PBSPro: job scheduling software**
  - **ClusterWorx: System management software**
  - **Various MPI debugging tools, libraries, performance profiling tools, and utilities**
  - **TotalView: FORTRAN/MPI debugging tool**
  - **PGI's Cluster Development kit, which includes parallel FORTRAN, C, C++ compilers**
  - **Intel Compiler**
  - **VMware ESX server**
- **Integration, configuration and optimization of this middleware with the NWP ops run has required a considerable amount of time and effort**

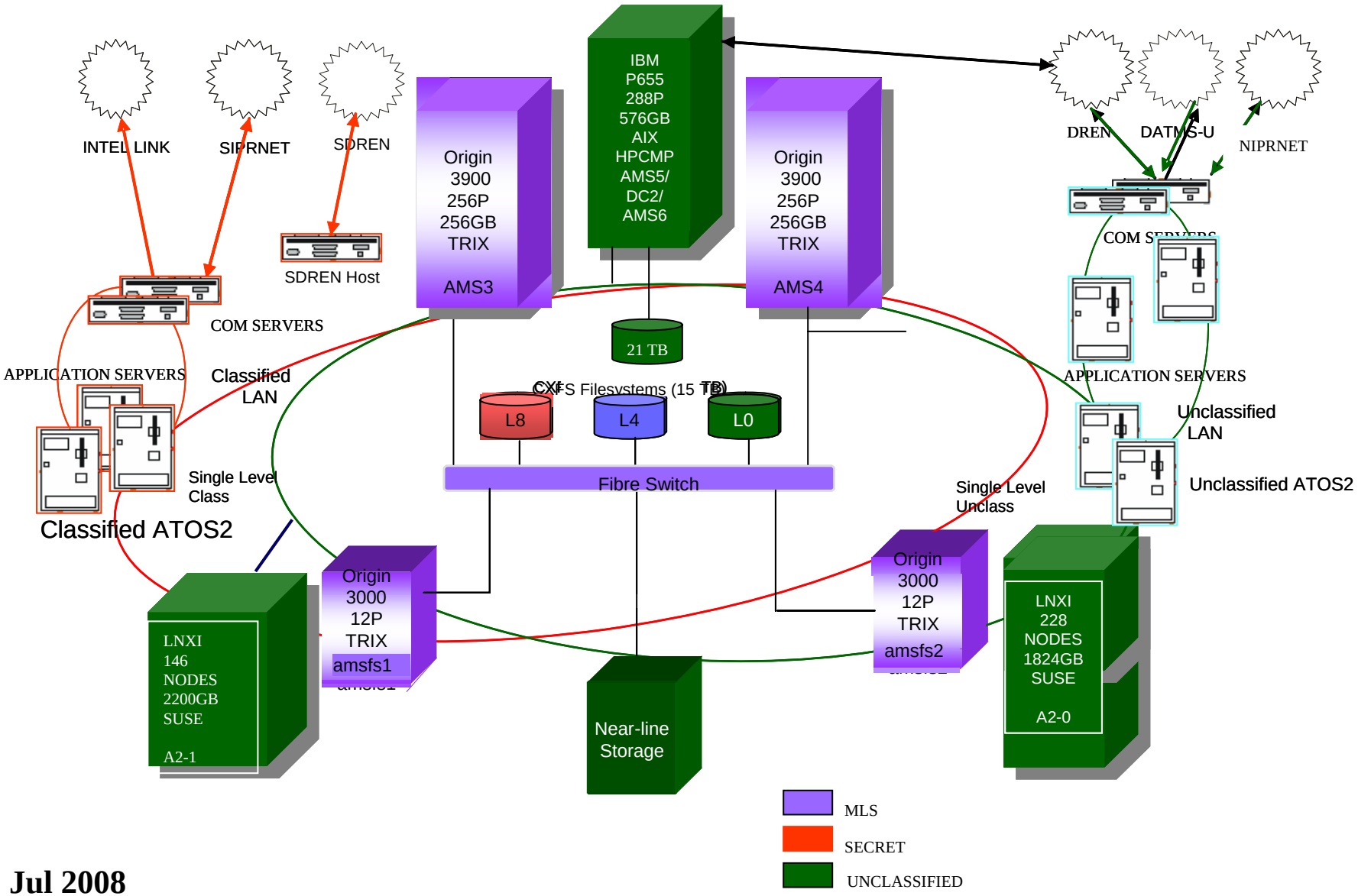
Other Challenges in Bringing A2 to IOC

- **Meeting required IA constraints**
  - **Hardening nodes on Infiniband networks to generic IA standards**
  - **Validating software stacks for security compliance**
- **Establishing required network connections, real-time data feeds, and operational databases for the new platform**
- **Porting model codes and other required applications to the new architecture and Operating System (OS)**
- **Implementing software Configuration Management (CM) procedures in the new environment**
- **Dealing with compiler and MPI library issues**
- **Optimizing I/O performance to achieve required model throughput**
  - **Disk reconfiguration**
  - **Systems software changes (e.g., GPFS)**
  - **Model software changes (e.g., COAMPS-OS)**
- **Establishing operational job scheduling and preemption with new job scheduler (PBSPro)**
- **Dealing with fallout from the LinuxNetworx business shutdown**
  - **Loss of expertise, integration/validation support and A2 test platform previously provided by LinuxNetworx**
  - **Engagement with third-party vendors previously subcontracted by Linux Networx**

A2 High-Level Timeline

- **A2-0 (Opal)**
  - –Procurement started Jun 06
  - –Full system on-site Jan 07
  - –IOC achieved Oct 08
- **A2-1 (Ruby)**
  - –Procurement started May 07
  - –System on site Sep 07
  - –Expect to achieve IOC Jan 09
- **A2-1 (Topaz)**
  - –Procurement started May 08
  - –System on site Sep 08
  - –Expect to achieve IOC Mar 09

POPS System Architecture



Jul 2008

POPS Computer Systems

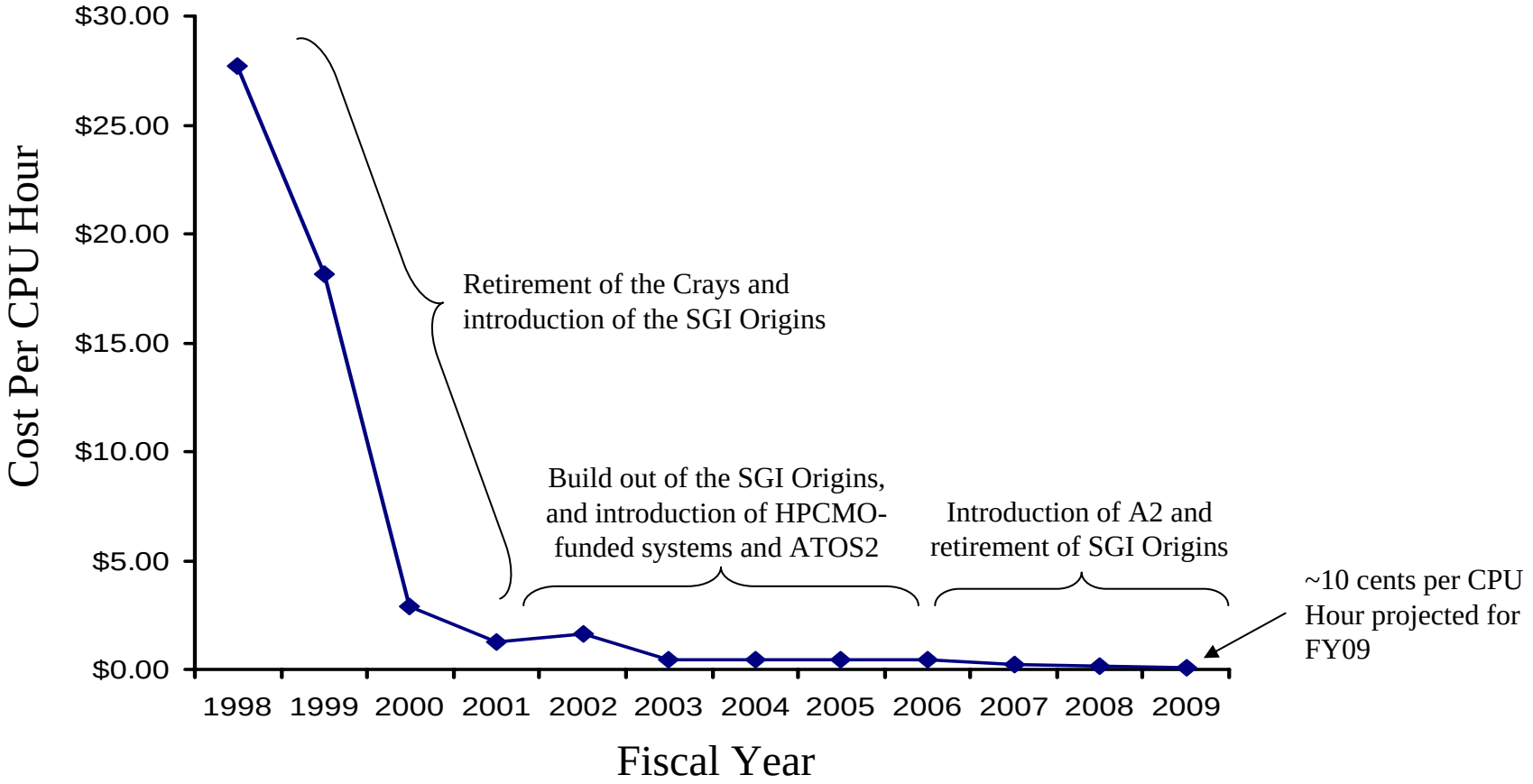
NAME	TYPE	#CPUs	MEMORY (GB)	PEAK SPEED (TFLOPS)	OS
AMS2	SGI ORIGIN 3800	128	128	0.2	TRIX
AMS3	SGI ORIGIN 3900	256	256	0.7	TRIX
AMS4	SGI ORIGIN 3900	256	256	0.7	TRIX
AMS5	IBM p655+	288	567	1.8	AIX
DC3	IBM p655+	256	512	1.9	AIX
ATOS2	IBM 1350s/x440s/x345s	438	645	2.2	Linux
CAAPS	IBM e1350s	148	272	0.7	Linux
A2-0	Linux Cluster	928	1800	10.0	Linux
A2-1	Linux Cluster	1168	1170	12.0	Linux
A2-2	Linux Cluster	126	1200	3.0	Linux
TOTAL		**3920**	**7837**	**~33**	

- **The POPS computer systems are linked directly to ~150 TB of disk space and ~160 TB of tape archive space.**
- **The POPS 30 TFLOP total is comparable to the 15% MSRC allotment for FY09 (~35 TFLOPS)**

Direct Engagement with HPCMO

- **FNMOC has competed successfully for three High Performance Computing Modernization Office (HPCMO) Dedicated HPC Project Investment (DHPI) systems:**
  - **2002 256 Processor SGI Origin 3900 (now designated AMS3)**
  - **2004 96 Processor IBM p655+ (now designated AMS5)**
  - **2006 256 Processor IBM p655+ (now designated DC3)**
- **Each DHPI system represents a $2-3M HPCMO investment that is free to Navy**
- **Following completion of 2-3 year DHPI project, typically involving FNMOC, AFWA and NRL collaboration, the systems are folded fully into POPS and used operationally**

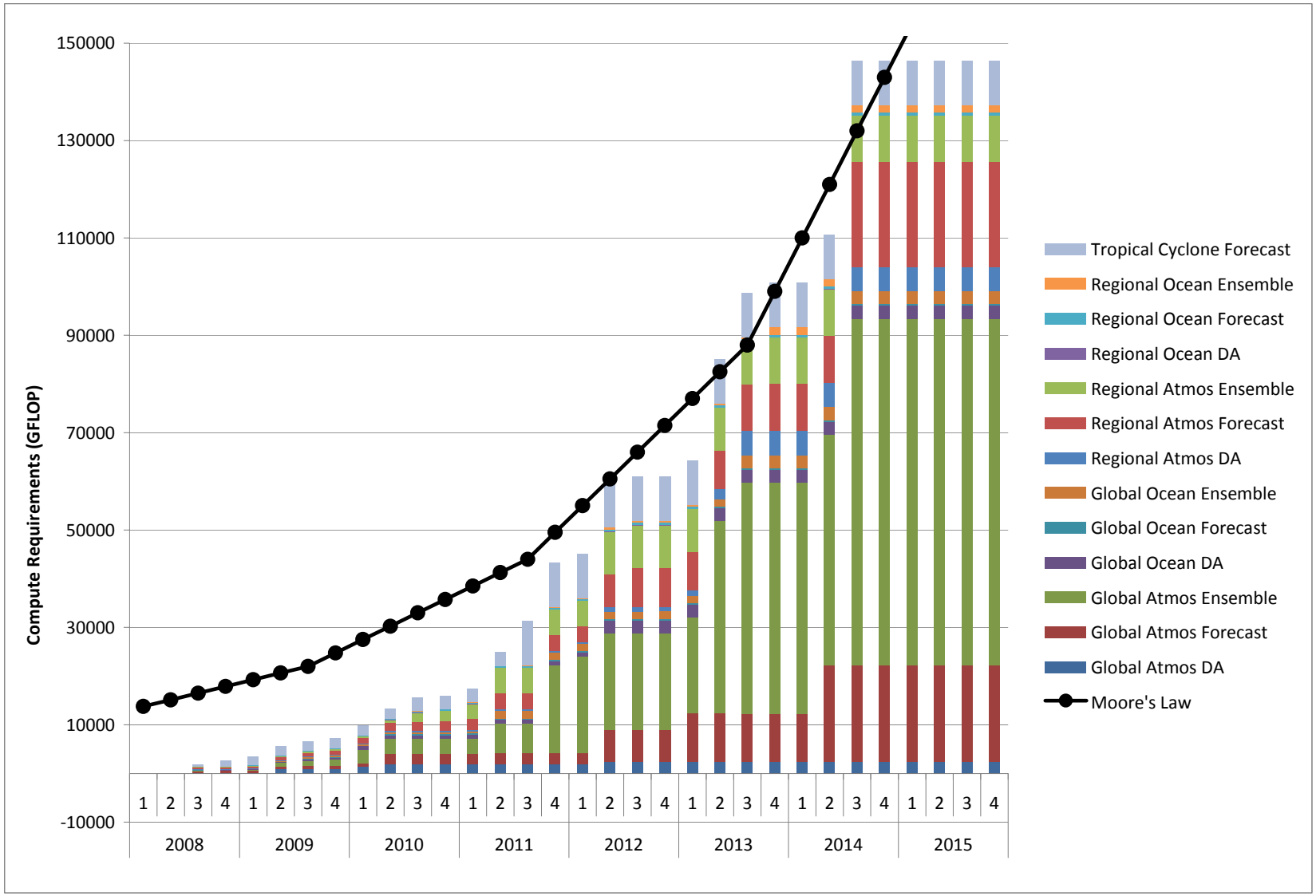
History of POPS Cost Per CPU Hour



Cost Per CPU Hour is the number of HPC CPU hours used in a given year by the total cost of ownership of the FNMOC HPC systems for that year. Yearly total cost of ownership includes the initial hardware and software procurement costs, prorated over the life of each HPC system, plus yearly contract costs for hardware and software maintenance and support.

Future (to 2025) HPC Outlook

Models and Data Assimilation Drive HPC Resource Requirements



Ensemble Modeling

- **Estimated HPC resources devoted to ensemble modeling (2015) - ~60%**
- **Current modeling initiatives dealing with ensembles**
  - **National Unified Operational Prediction Capability (NUOPC)**
    - Global ensemble modeling initiative
    - Participants: FNMOC, AFWA, NOAA
  - **Interagency Environmental Modeling Concept of Operations**
    - Mesoscale ensemble modeling effort
      - **Domain alignment (FNMOC & AFWA)**
      - **Exchange of model data**
      - **Ensemble members initialized by global ensemble members**
  - **Joint Ensemble Forecast System (JEFS)**
    - Mesoscale ensemble modeling effort (current)
    - FNMOC & AFWA

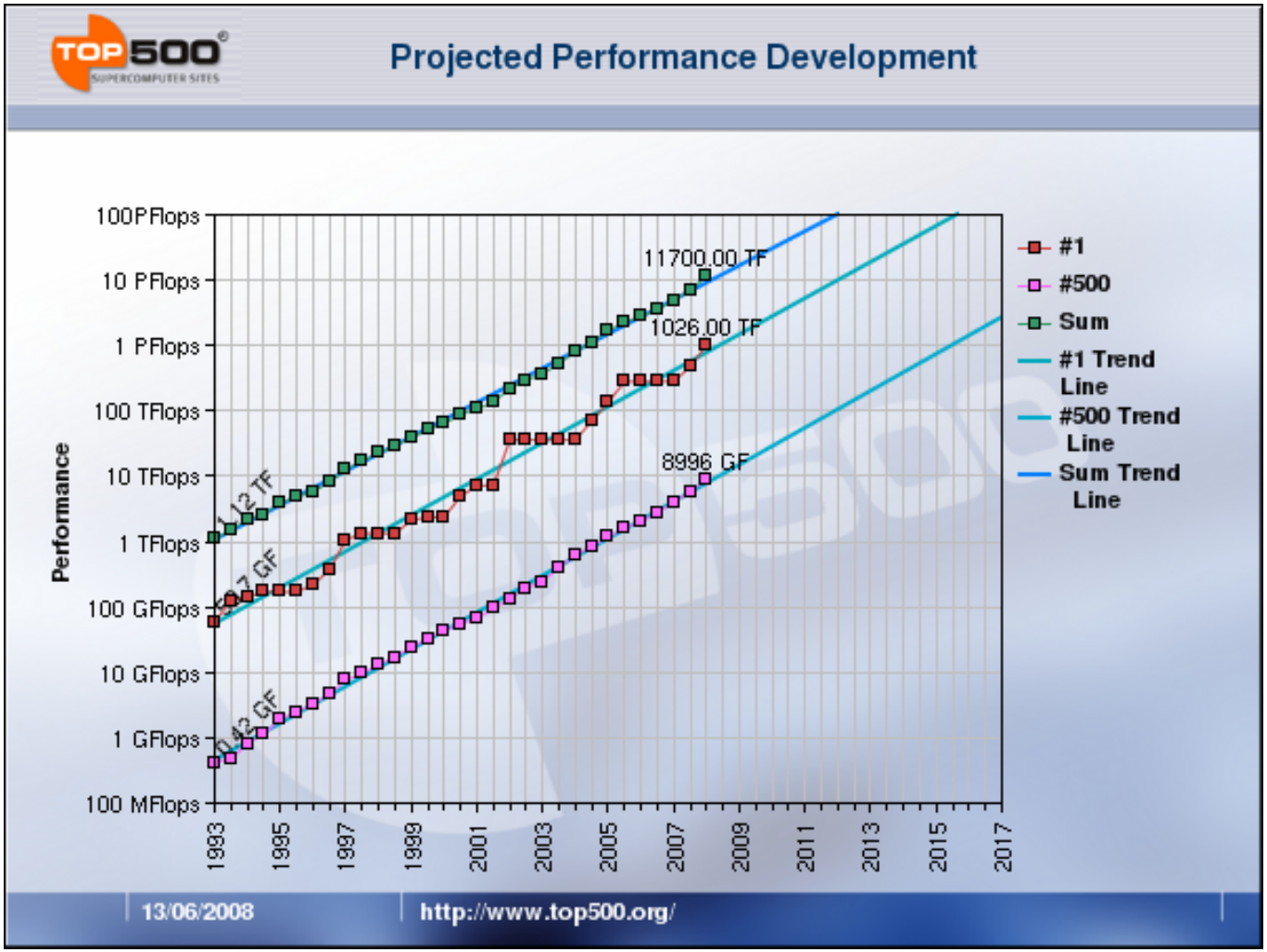
Data Exchange

- **Information Assurance (IA) Issues**
- **Current modeling initiatives dealing with ensembles**
  - **Interagency Environmental Modeling Concept of Operations (IEMCO)**
    - Committee for Operational Processing Centers (COPC)
    - Signatories: AFWA, FNMOC, NAVO, NCEP
    - Global and mesoscale ensemble modeling efforts
  - **National Unified Operational Prediction Capability (NUOPC)**
    - Global ensemble modeling initiative
    - Participants: FNMOC, AFWA, NOAA
  - **Joint Ensemble Forecast System (JEFS)**
    - Began as “Proof of Concept” for ensembles
    - AFWA/FNMOC joint initiative

IA Guidance Documents

- **AFWA:**
  - **DoD Instruction No. 8500.2, "Information Assurance (IA) Implementation"**
  - **Air Force Instruction 33-202, Vol. 1, "Network and Computer Security"**
  - **Briefing, "Certification & Accreditation", space information assurance**
- **FNMOC:** ***(not for distribution)***
  - **"Navy-Marine Corps Unclassified Trusted Network Protection Policy", v1**
  - **"Navy UTN*Protect* Policy, Sec. 4.2, update, 1June 2006**
- **NCEP:**
  - **"Information Assurance for NCEP's Numerical Weather Forecast Systems, S. Lord, 4 April 2008. (preliminary statement)**

Projected Performance of TOP500 HPC Systems to 2017



If projected rate of performance increase continues, would expect mid-range TOP500 systems to have performance levels between about 1,000 and 10,000 PFLOPS in 2025.

Moore's Law and Model Resolution

- **Thus, if continue to grow Naval Oceanography Program HPC resources in proportion to Moore's Law through 2025, model grid spacing enabled in 2025 would be about:**

$$2^{-(2025-2008)/8} \sim 1/4 \text{ current values}$$

- **Issues:**
  - **–R&D required in model dynamics and physics to support these resolution increases**
  - **–Real-time and other supporting data required to support these resolution increases**
  - **–Higher relative proportion of available cycles going to more sophisticated data assimilation (e.g., 4DVAR, Kalman Filter, etc.)**
  - **–Tradeoff of model resolution versus number of ensemble members, as modeling focus shifts increasingly toward ensembles**

Future of Moore' Law

- **Though its demise has been predicted many times, Moore's Law has essentially held for over 40 years; the semiconductor industry has repeatedly found new ways to sustain it.**
- **But even Gordon Moore has doubts; recent quote: "Moore's Law should continue for about another decade. That's about as far as I can see."**
- **Current technology roadblock looming: behavior of semiconductors begins to be affected by quantum mechanical effects at about 10 nm, which will limit processor components to about 15 nm size.**
- **New technologies that may likely be employed to sustain Moore's Law even beyond this physical limitation:**
  - **transition from 2D to 3D chips, with as many as 20 layers of circuitry in a vertically integrated stack**
  - **Use of carbon nanotubes for chip channels**
  - **optical computing (i.e., using photons rather than electrons for switching and communications on the chip)**
  - **New electronic components (e.g., the "memristor") that can replace and be fabricated much smaller than transistors**

Likely HPC Technology Trends Between Now and 2025

- Continued development of faster processor cores based on technologies like those listed on the previous slide
- More cores per socket and more sockets per motherboard (e.g., expect 16 cores per socket and 8 sockets per motherboard by 2011)
- Improvements in disk I/O
  - Solid-State Disks
  - Focus on small file I/O and metadata
- Dynamically reconfigurable computing environments
- Algorithm optimization at the hardware level
  - Field Programmable Gate Arrays (FPGAs)
  - New programming languages and tools for hardware-level code optimization
- Faster interconnects and communications, possibly based on Laser technology
- Solid state storage at the molecular level replacing rotating persistent storage
- Improved programming languages for HPC