The NCEP HPC Workload 2010 State and Forward Challenges

George VandenBerghe
I.M. Systems Group.
November 2010

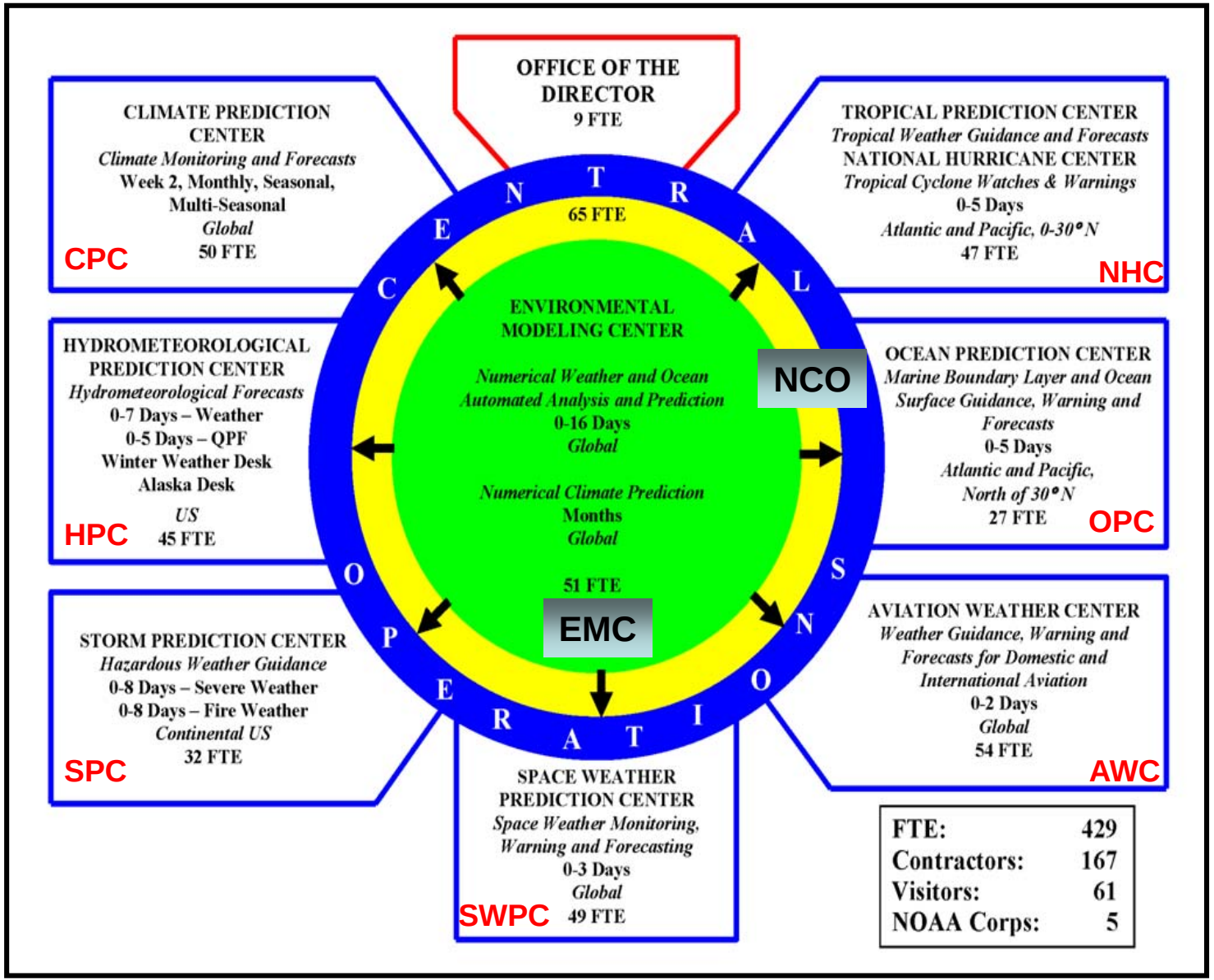
Overview

- NCEP SITE
- NCEP FUNCTION
- NCEP COMPUTERS
- NCEP WORKLOAD
- COMPUTING PROPERTIES
- WORKLOAD TRENDS
- SUMMARY

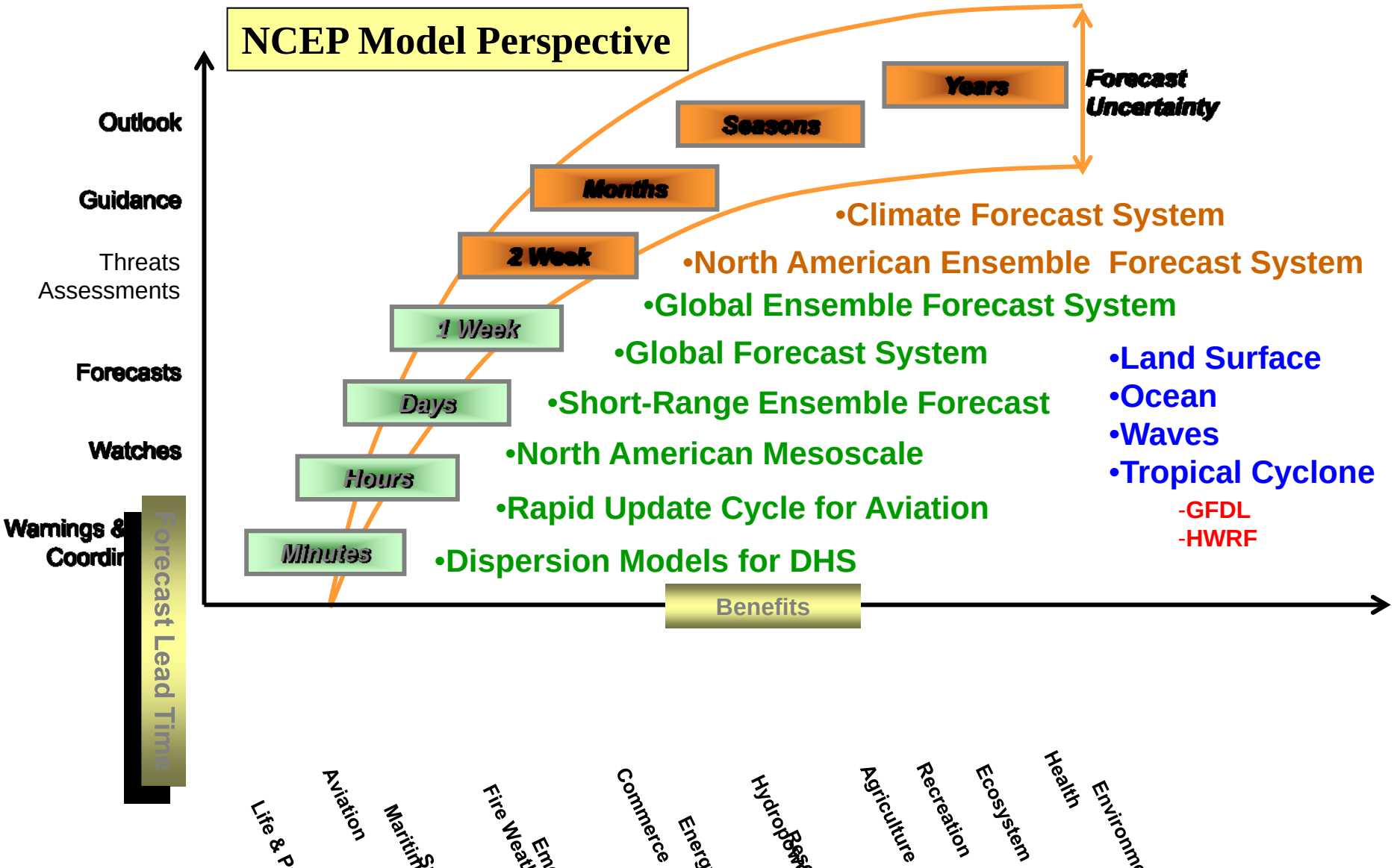
NCEP's Work

- National Center for Environmental Prediction (NCEP)
- Primary source of weather forecast guidance for U.S and substantial source for international users as well.
- Private companies and media start with guidance and tailor it for customer and audience needs providing significant added value.
- MOST GO TO NCEP FOR BASIC DATA
- Differs from ECMWF in that ECMWF does medium range and longer timescale and NCEP does all timescales.
- Forecast process requires MUCH computer power!!

The National Centers for Environmental Prediction



NWS Seamless Suite of Forecast Products Spanning Weather and Climate



Past Platforms

- ENIAC 1 kflop
- IBM 70x mid 50s 10 kflop
- IBM709x early 60s 100 kflop
- CDC6600 mid 60s-early 70s: 1 mflop
- IBM 360/195 x3 70s-early 80s 10 Mflop
- CDC CYBER205 x2 80s 100 Mflop
- CRAY Y-MP8 early 90 1Gflop (1.3 aggregate)
- CRAY C90 mid to late 90s 6 Gflop (8 aggregate)
- IDM SP 1999-2000 30 gflop (70 aggregate)
- IBM SP 2000-2002 60 gflop (140 aggregate)(x2)
- IBM P690 2003-2004 160 gflop (320 aggregate)(x2)
- IBM P655 (1000 aggregate)(x2)
- IBM P5 (2800 aggregate)(x2)
- IBM P6 (9000 aggregate) (x2)
- ***Available cycles have doubled every two years since 1950 with***
- ***Faster increase in recent years.***

Supercomputing at NCEP

IBM Power6 p575

69.7 Teraflops Linpack

156 Power6 32-way Nodes

4,992 processors @ 4.7GHz

19,712 GB memory

320 TB of disk space per system

13 PB tape archive (18PB Cap.)

Fairmont, West Virginia

Cirrus— (backup)

Gaithersburg, Maryland

Stratus— (primary)

Current NCEP CCS 10/2010

- Two IBM Power6 clusters.
- 144 compute nodes each with 32 dual core SMP cpus.
- 128GB of memory per node.
- 320 TB of GPFS on each cluster.
- 320 Additional TB will soon be added on each cluster
- 150TB of additional shared GPFS will soon be added.
- 10Gb WAN between clusters.
- 18PB HPSS archive in the Gaithersburg MD facility, accessed *(successfully with full functionality)* through WAN at Fairmont WV.
- (HPSS rates are 30mb/sec from Fairmont and 40-50 from Gaithersburg (100/node from both with aggregate transfers)

NCEP COMPUTING

- There have been ~~28~~ ~~31~~ 33 doublings in compute capacity since 1954.
- The crossed out figures were from the Summer 2002 and 2006 Spscicomp presentations.
- Long term site is following “Moore’s Trend”
- Processor counts, flat for six years, have now doubled three times since 2003.

WORKLOAD OVERVIEW

- Workload divides to DEV and PROD.
- PROD has priority but usually runs on own machine. DEV runs on the other machine.
- Prod mirrors its filesystem state to the Dev machine with rdist. (this is reaching scalability limits)
- DEV and PROD can switch machines in ~10 minutes. (all active jobs lost)
- Switches done for maintenance, testing, and disaster mitigation (rare!)

INFASTRUCTURE IMPROVEMENTS

- 1990 NO UPS MTTI 300 hr.. 2 hr repair
- 1993 DISK UPS DISK MTTI 10x increase
- 1995 RAID eliminates disk hardware losses (well almost)
- 1995 CRAY J UPS. 7x24 ops possible
- 1999 IBM UPS 7x24 undegraded ops possible
- 1990 5% vendor down +2-4% site out
- 1999 <1% down + 0.1% site out
- 2004 Geographic separation of dev and prod systems reducing region wide event risk. (one is in Washington DC suburbs and the other is in West Virginia 300km distant.)
- 2010 (pending) Multicluster GPFS across WAN.

WORKLOAD

Workload consists of many independent threads.

Regional (limited area) modeling, Global modeling, Hurricane forecasting, short term climate state prediction (coupled models), ocean modeling (hycom), wave forecasting, air quality modeling, Global ensembles, regional ensembles, Regional nested models, Global (now regional also) data assimilation (the NCEP GSI), post processors for all of this, and product generators for all of this.

Mpi apps scale to hundreds (GFS, GSI) or low to mid thousands of tasks (NMM, HYCOM, WAVE). NMM nest may scale to more Threading gets GFS scalabiltiy to mid thousands of cores

- MPI comm performance constrained by switch bandwidth.
- MPI comms are a small (15-20% of total time) fraction of total time.
- Computing dominates our work, not comms.
- Much more threading than in 2002 when I last presented this. Both NCEP and IBM have improved threading, and its API considerably.
- Combination of limited MPI scalability and very good threading requires a very good Open_MP implementation.

We scale enough but don’t scale well

- Not a single NCEP forecast or analysis app is constrained significantly by scalability on our 4608 processor clusters. (what??)
- This is because we run many concurrent work threads and ensembles. Dividing 4608 by “many” yields only a few dozen to hundred processors available for a problem and less for ensemble members.
- This will change as processor counts increase and with a slight shift towards deterministic forecasts.
- But points 1 and 2 make an argument that we need to look at scalability, more difficult to support and easier to put off.

MAJOR APPS

- Global Spectral Model (GFS) T574L64
- 12KM Regional model run to 84H
- GSI analysis for both of these. Work with ENKF and 4DVAR is ongoing.
- Nested hurricane model (GFDL and HWRF)
- Quick turnaround regional grid model (RUC)
- Ensemble forecasts (Regional and Global)
- Coupled GFS/Ocean model (for climate forecast)
- Wave, Air Quality, Hycom models.
- Swarms of small (and big) pre and post processors

SCHEDULING

- PROD has highest priority.
- PRODS must run same time daily.
- **EARLY completion also causes problems.**
- High variance/low runtime not as desirable as low variance/longer runtime. Numerous jobs run together.
- Main constraint is CPU speed. memory not as constraining (this makes schedule planning easier). I/O is also now a significant and rapidly worsening constraint. Theoretical switch contention issues have not appeared in practice.

Current Workflow

- One job runs a forecast model (e.g. GFS)
- Each 3 hour forecast write triggers a post processor job which runs a few MPI tasks on a fraction of a node.
- Output from these post processors is read by product generators. Products (standard Grib fields, graphics) are disseminated externally.

Workflow

- The forecast job (or ensemble members) run on parallel nodes (all but four)
- The post processors run on four “Prodser” nodes intended for smaller jobs.

Serial nodes also do prod file mirroring. (turning this into a daemon is on our list of considerations but is not a high priority)

I/O has been a bottleneck on individual serial nodes.

(If we spread serial work across a larger number of nodes we expect a much harder to mitigate filesystem bandwidth problem in a few years)

Parallel node jobs mostly run not_shared. (don’t share nodes)

“Prodser” jobs run shared.

We have a LOT of memory per node and per core and it does not constrain us much when scheduling.

Near term Evolution

- Post processor I/O is a bottleneck.
- Plans are to gradually incorporate post processors into the major model executables using a few additional nodes.
- Post processor(s) and forecast model run as one MPI executable.
- If all intermediate states are still saved this reduces I/O by almost 2X.
- If some intermediate states are discarded I/O reduction is (perhaps much) larger.

Archive Problem!!

- With larger HPC platforms and desire for more and more ensembles, our data volume is exploding (scales with compute capacity)
- I/O technology trends are not matching this. (physical density is, cost and performance aren't)
- We cannot afford an archive that scales with compute.. Starting NOW

Archive Drivers.

- Archive data is of three classes or “Types”
- Type I. Production model forecast archive. (scales with production capacity)
- Type II User general data (including pack rats) (scales with development cpu capacity)
- Type III. Major dataset production (Reanalysis and Reforecasts). Upper bound scales with development cpu capacity but with a much larger coefficient. Constrained by management policy
- Everyone goes after type II but types I and III are the actual drivers.

Breakdown

- Type I 6PB
- Type II 3PB
- Type III 4.5PB (exploded from 1PB in 3Q 2007 when a major reanalysis was done)

- Types I and III are more easily managed. Type I is determined by decisions on what to save and how to replicate what isn't
- Type III is determined by major science initiatives and dataset size is known at decision time.
- However we've only postponed issues with type II.

Archive Problem.

- Development cpu capacity is expected to increase by 4-6X.
- Archive size MIGHT increase by 2x, probably less.
- No major type III initiatives are planned until FY2014, (I'm assuming one however) then we have a crunch with a big one likely 2014-2016.
- Type I base growth will be low for next few years (every cloud has a silver lining??)
- Discussions on how to handle this are not yet public (so I just state we have a problem)

Scaling Problems.

- GFS has 1D decomposition with excellent threading.
- This is fine for now but needs reexamination on future platforms.
- Scaling is ¾ wavenumber x threadcount.
- At NCEP for T574 with 16 threads this can use 6000 cores.
- 2D decomposition might ideally scale to 10000+ MPI tasks (see my 2006 presentation (http://www.ecmwf.int/newsevents/meetings/workshops/2006/high_performance_computing-12th/pdf/George_Vandenberghe.pdf)) but physics load imbalance will reduce that, perhaps a lot.

Scaling Problems.

- Grid point, finite difference Eulerian forecast models should scale to 100,000 or more MPI tasks *if load imbalance is not considered*. (Semi Lagrangian won't do as well{ George Mozdzynski. talk presented at this conference suggest this may be an even bigger problem})
- *But load imbalance is a big problem and we have to mitigate it.*
- We don't have this issue yet because we have "only" 4608 cpus per cluster.
- Author's (GWV) belief is that above problems will be tractable if we get large corecount machines and are motivated to address them.
- Scaling of analysis is a tougher issue but GSI scales enough (600 cores) for now and ENKF is constrained by compute costs, not just scalability.
- It's hard to look at scalability when constrained by capacity.

I/O

- Compute metrics increasing faster than disk metrics.
- 63x degradation in disk bandwidth/flop since 1982
- 100x degradation in aggregate tape bandwidth/flop since 1982.
- People focus on space but normalized bandwidth is getting worse faster
- (2-4x space/flop decline, 60-100x bandwidth/flop decline since 1982.. Slow but exponential decline.)
- Space/flop can be mitigated with budget adustments.
- *It is author (GWV) opinion that bandwidth/flop cannot be mitigated in HPC with just more money and will require computing paradigm adjustments*.

Memory

- We have 128GB/node (2GB/SMTcpu) at NCEP.
- Large memory apps are common.
- Memory cost and power will drive memory/core down on future machines.
- Avoid full domain operations on one core (and eventually on one node) anywhere.
- Memory use/task will have to approach the ideal of scaling inversely with taskcount...

Summary

- NCEP is cpu CAPACITY bound.
- Capacity constraints conceal a scalability issue.
- I/O will be an increasingly intrusive issue on future platforms.

- Questions??