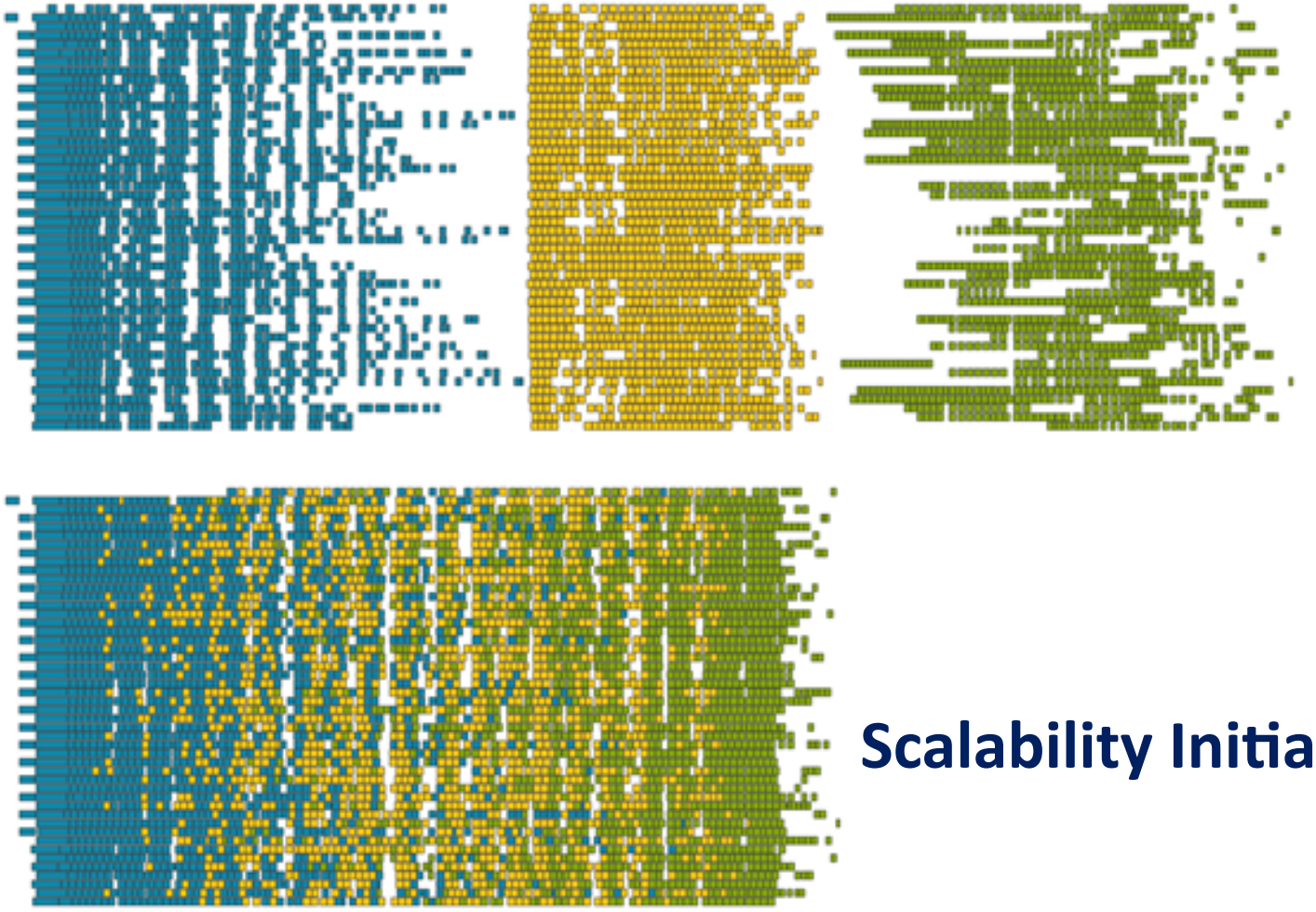


Picture: Stan Tomov, ICL, University of Tennessee, Knoxville

Scalability Initiative at ECMWF

Peter Bauer,
Mike Hawkins, George Mozdzynski,
Deborah Salmond, Stephan Siemen,
Peter Towers, Yannick Trémolet,
and Nils Wedi

Weather vs climate prediction

	Weather	Reanalysis	Climate
Resolution/time step:	15 km, L137 (0.01 hPa), 10' (ensembles = ½ high-resolution)	80 km, L60, (1 hPa), 15'	80 km L199 (0.01 hPa), 2'
Time constraint:	10 d/h = 240 d/d		8 m/d = 240 d/d (→ 10 y/d = 3650 d/d)
Prognostic variables:	p_s, u, v, T, q, $q_{l/i/r/s}$, cc	= weather	= weather + composition
Coupling:	none (ocean soon) (ensembles: ocean, sea-ice soon)	none (ocean soon)	ocean, sea-ice
Data assimilation:	atmosphere, surface (uncoupled)	= weather	surface, atmosphere (coupled)
Model core:	hydrostatic, spectral	= weather	= weather
Critical physical paramerization:	radiation (= ½ others)	= weather	= weather
HPC cores:	O (10k)	O (0.1k)	O (1k)

In simplified terms:

→ Resolution etc.: climate = weather – 5-10 years

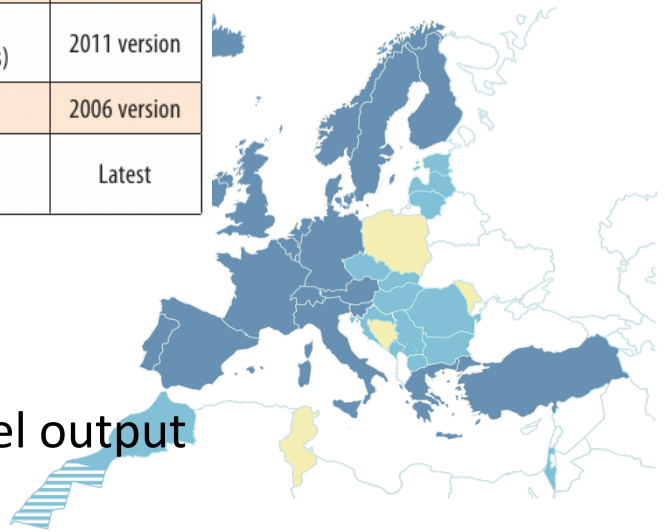
→ Earth system components etc.: weather = climate – 5-10 years

ECMWF

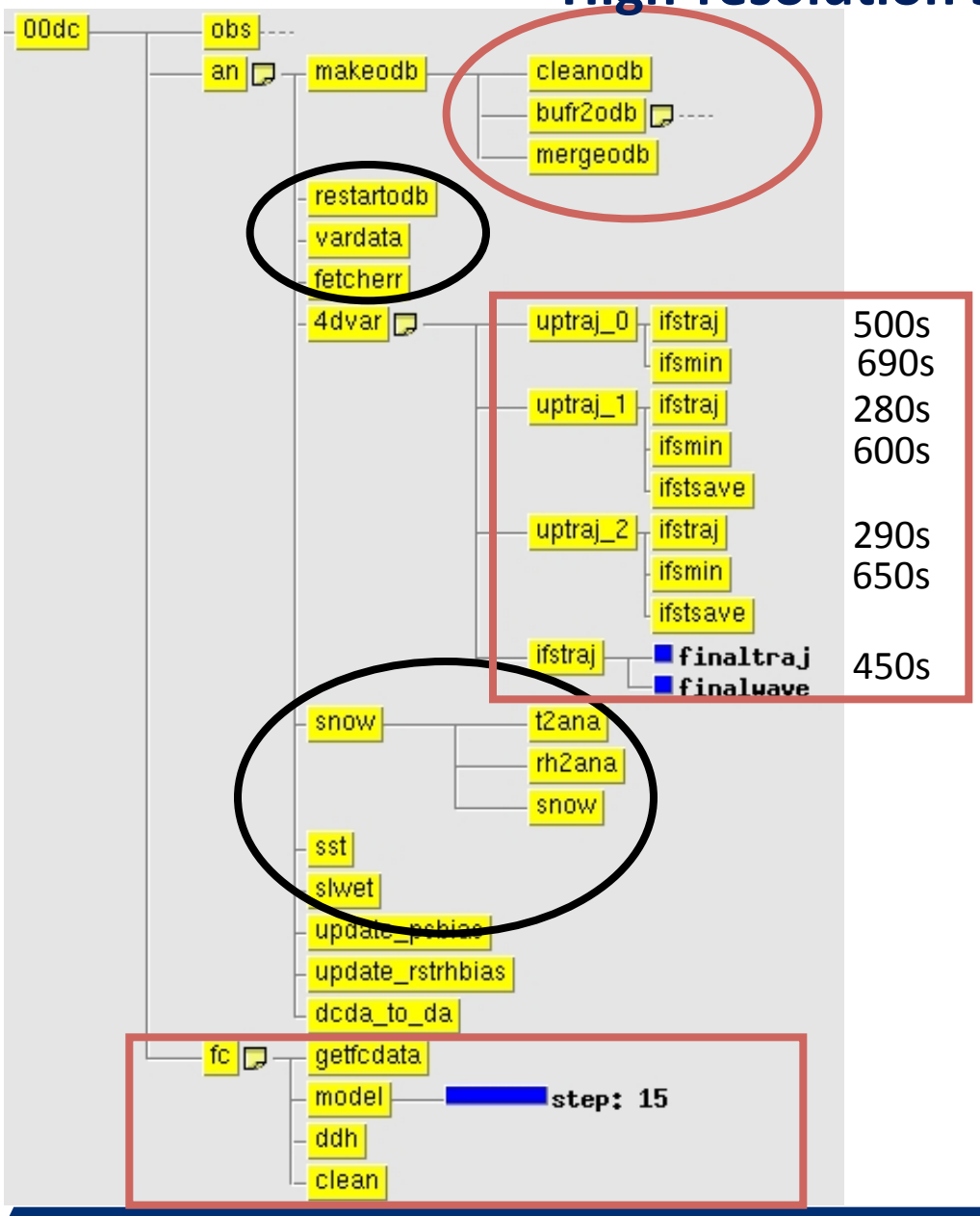
- An independent intergovernmental organisation; established in 1975
- 20 Member States, 14 Co-operating States, 45 M£ annual budget (staff:HPC)
- 270 staff (110 RD, 55 CD, 65 FD, 40 AD)

	Forecast/Analysis	Number of members	Horizontal resolution	Vertical levels and pressure at model top (hPa)	Perturbation models	IFS cycle
HRES	Forecast 0–10 days	1	T1279/16 km	L137/0.01	No	Latest
ENS	Forecast 0–10 days	51	T639/32 km	L91/0.01	Yes (in analysis and model physics)	Latest
	Forecast 10–32 days		T319/64 km			
4DVAR	Analysis	1	T1279/16 km (T255 inner loops)	L137/0.01	No	Latest
EDA	Analysis	11	T399/50 km (T159 inner loops)	L137/0.01	Yes (in observations and model physics)	Latest
SEAS	Forecast 0–13 months	51	T255/80 km	91/0.01	Yes (in analysis and model physics)	2011 version
ERA	Analysis	1	T255/80 km	60/0.1	No	2006 version
BC	Forecast 0–90 hours, hourly output	1	T1279/16 km	91/0.01	No	Latest

- No regional systems
- No warnings issued
- Only limited value adding applied to forecast model output



High-resolution suite



BUFR to ODB:
200 sec, 4x(8-16PEs)

Fetch background forecast:
275 sec, 2x(1PE)

Analysis trajectory, minimization and update:
3460 sec, (3072PEs)

Surface analysis:
720sec, (3072PEs)

10-day forecast 1440 t-steps:
3900 sec, (3072PEs)

ECMWF data processing

- **Observations per day:** **100 Gbyte**
  Observations need to be re-processed in 30 minutes should the database be lost. In a regular situation ca. 30-50 GByte need to be transferred and pre-processed in less than 20 minutes. Feedback slightly larger but no time constraint.

- **Model output per day:** **12 Tbyte**
  The elapsed time of the analysis is about 50 minutes, HRES takes 60 minutes, and the first 10 days of ENS less than 60 minutes. The total elapsed time of a main forecast cycle is about 3.5 hours, but hardly anything is written out during the running of the analysis (first 45 minutes).

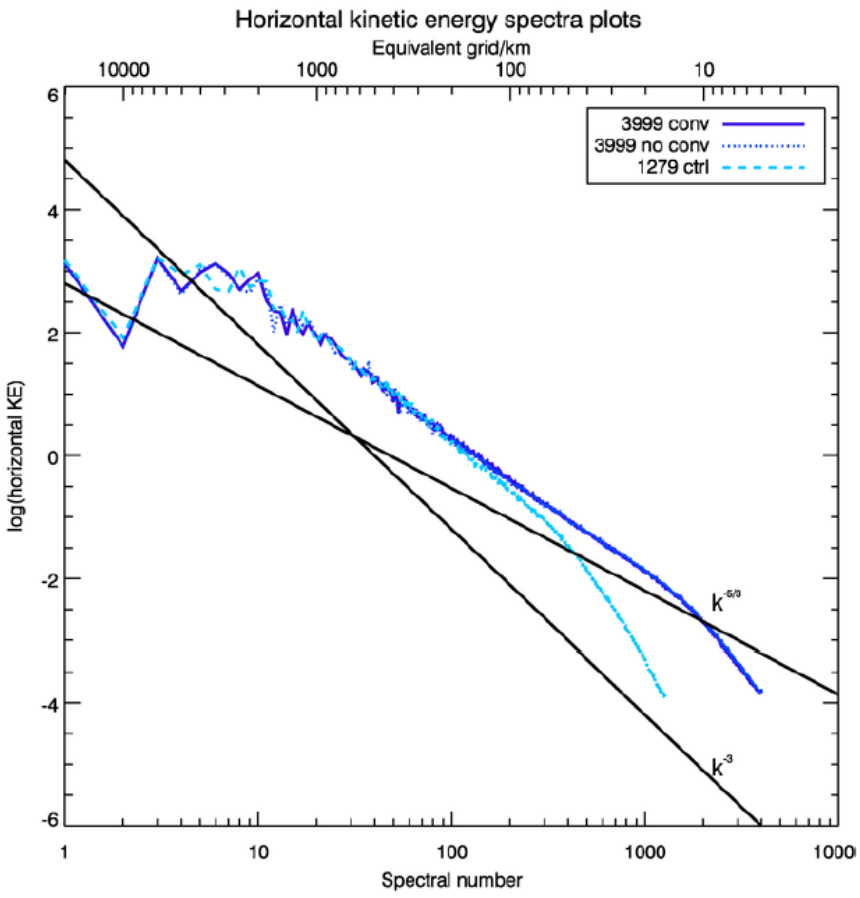
- **Products generated per day:** **6 Tbyte**
  Product generation needs to run alongside the model, i.e. in addition to writing the above model output. Production generation reads the data, processes it and writes it out again. The total elapsed time is identical, 3.5 (2.5) hours.


ECMWF saves **all** analysis input/feedback and model output (since day-1)!

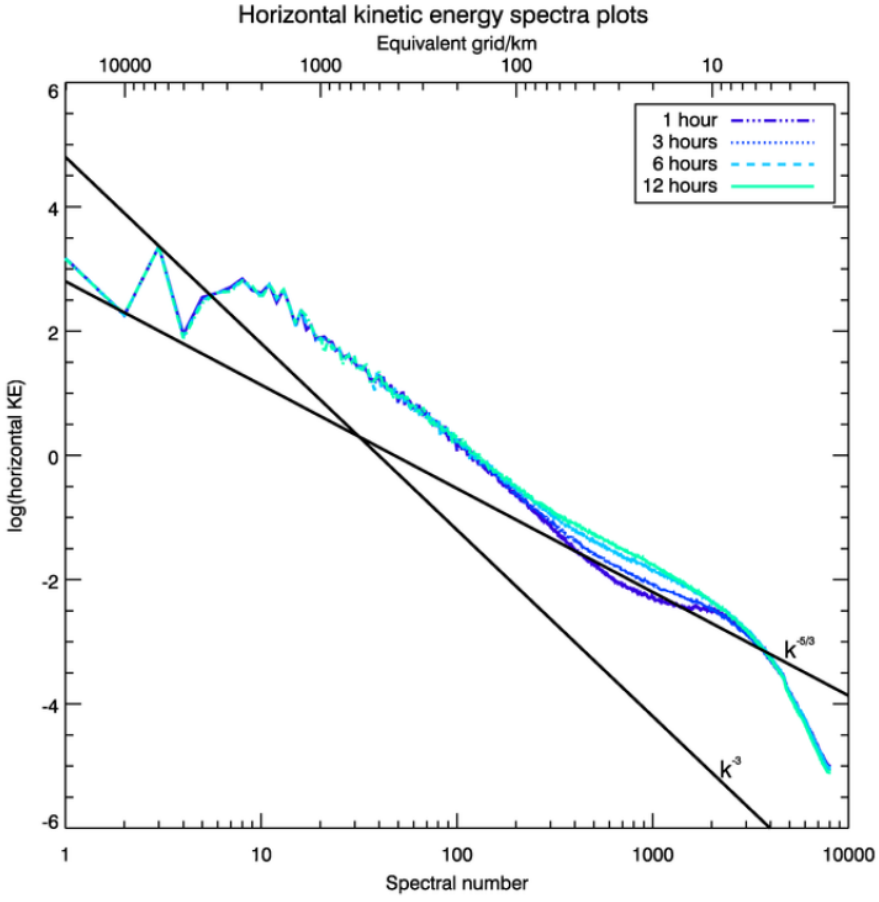
NWP: Benefit of high-resolution

500 hPa geopotential height energy spectrum from non-hydrostatic model integration

T1279/T3999 (10 days)

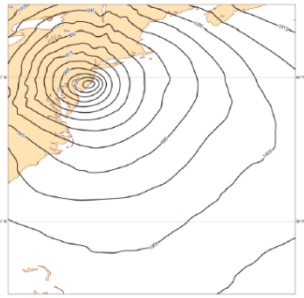


T7999 (1-12 hours)

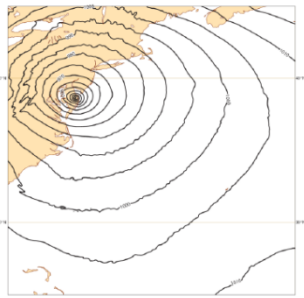


NWP: Benefit of high-resolution

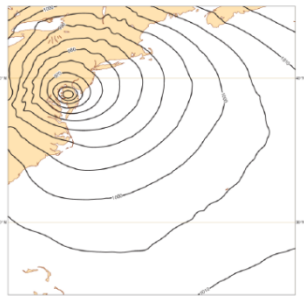
MSLP: AN 30 Oct



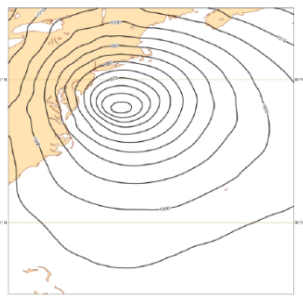
5d FC T3999



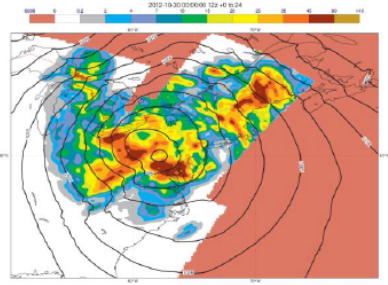
5d FC T1279



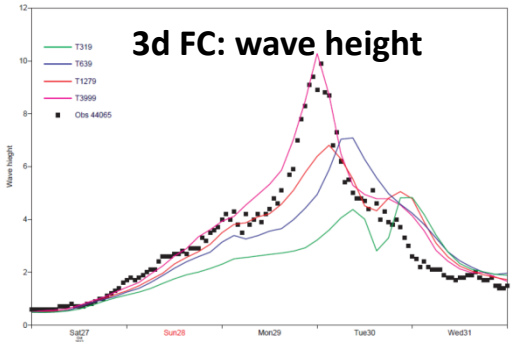
5d FC T639



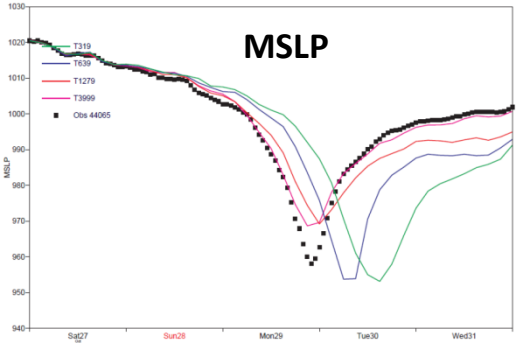
Precip: NEXRAD 27 Oct



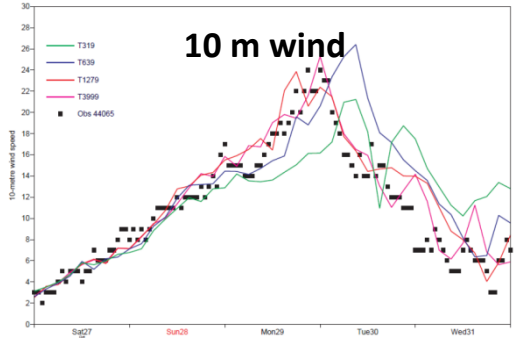
3d FC: wave height



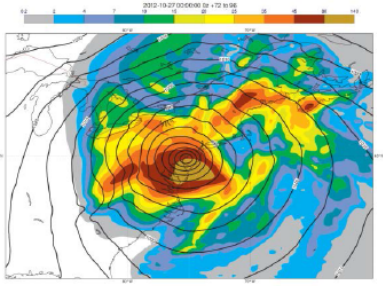
MSLP



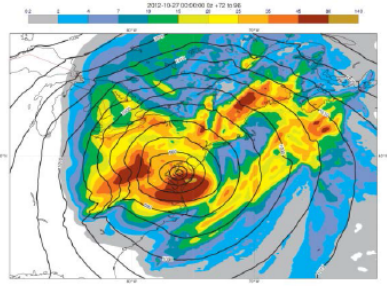
10 m wind



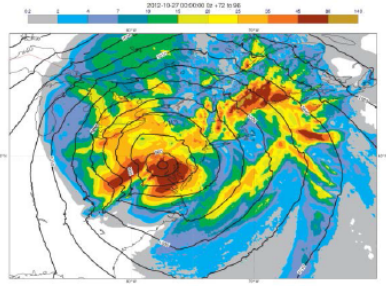
4d FC T639



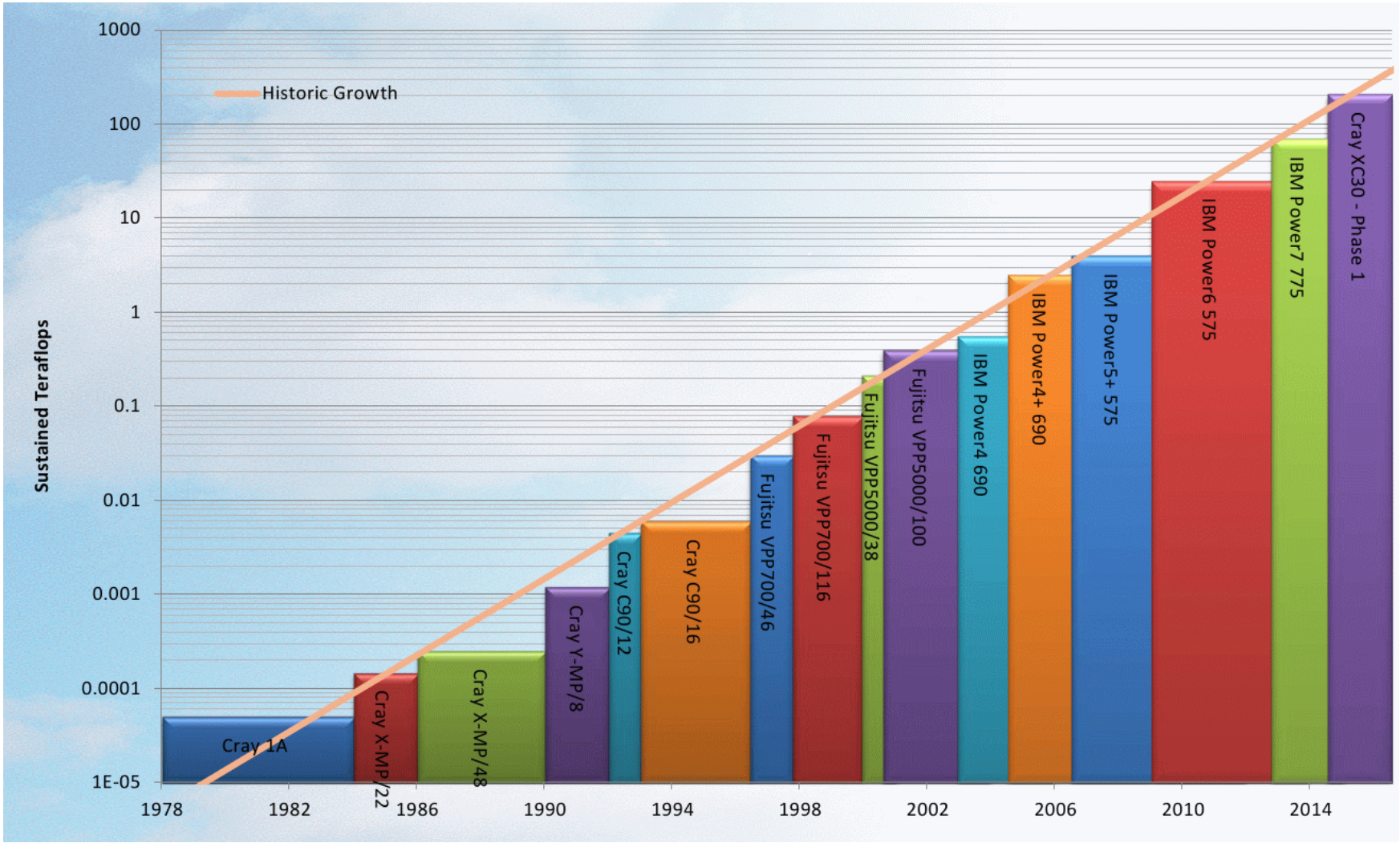
4d FC T1279



4d FC T3999



ECMWF HPC history

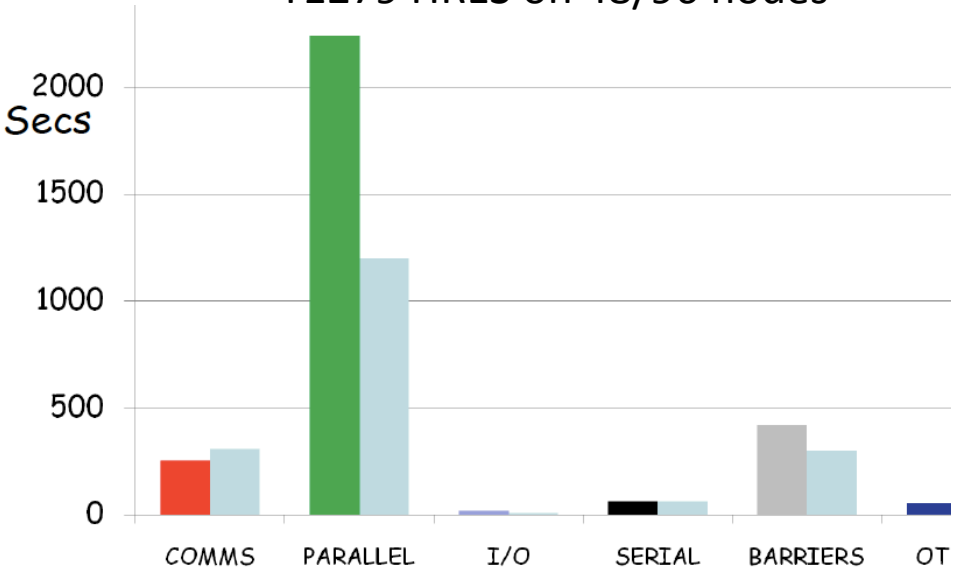


IBM P7 and Cray XC-30

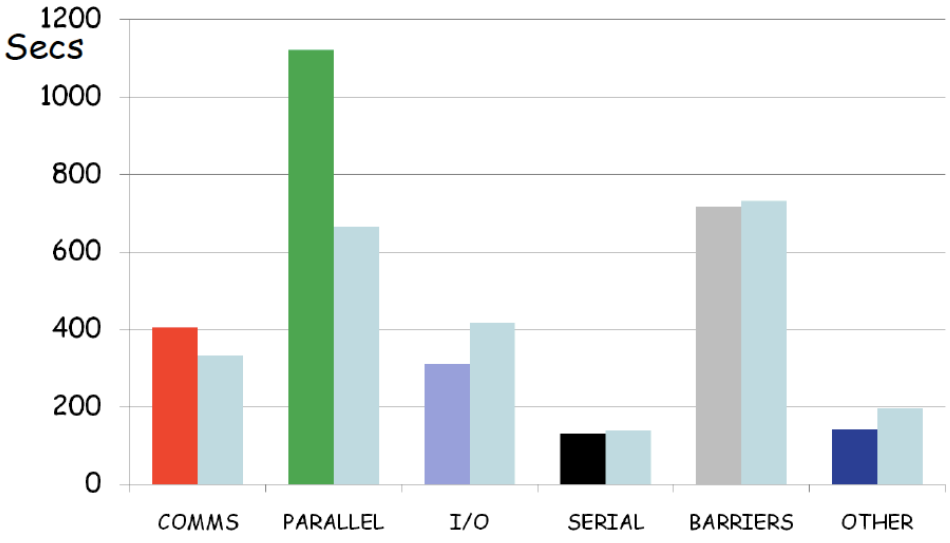
	Current	New
Sustained performance	~70 teraflops	~ 210 teraflops
Peak performance	~1500 teraflops	~3480 terfalops
Compute clusters	2	2
Each compute cluster		
Compute nodes	739	~3,500
Compute cores	23,648	~84,000
Total memory (TiB)	46	~210
Pre-/post-processing nodes	20	~64
Operating System	AIX 7.1	SUSE Linux/CLE
Scheduler	IBM LoadLeveler	Altair PBSpro/ALPS
Interconnect	IBM HFI	Cray Aries
Each storage system		
High performance storage (petabytes)	1.5	Over 3
Filesystem technology	GPFS	Lustre
General purpose storage (terabytes)	N/A	38
Filesystem technology	GPFS	NFS via NetApp FAS6240 filer

Experiments with IFS: Main components

T1279 HRES on 48/96 nodes

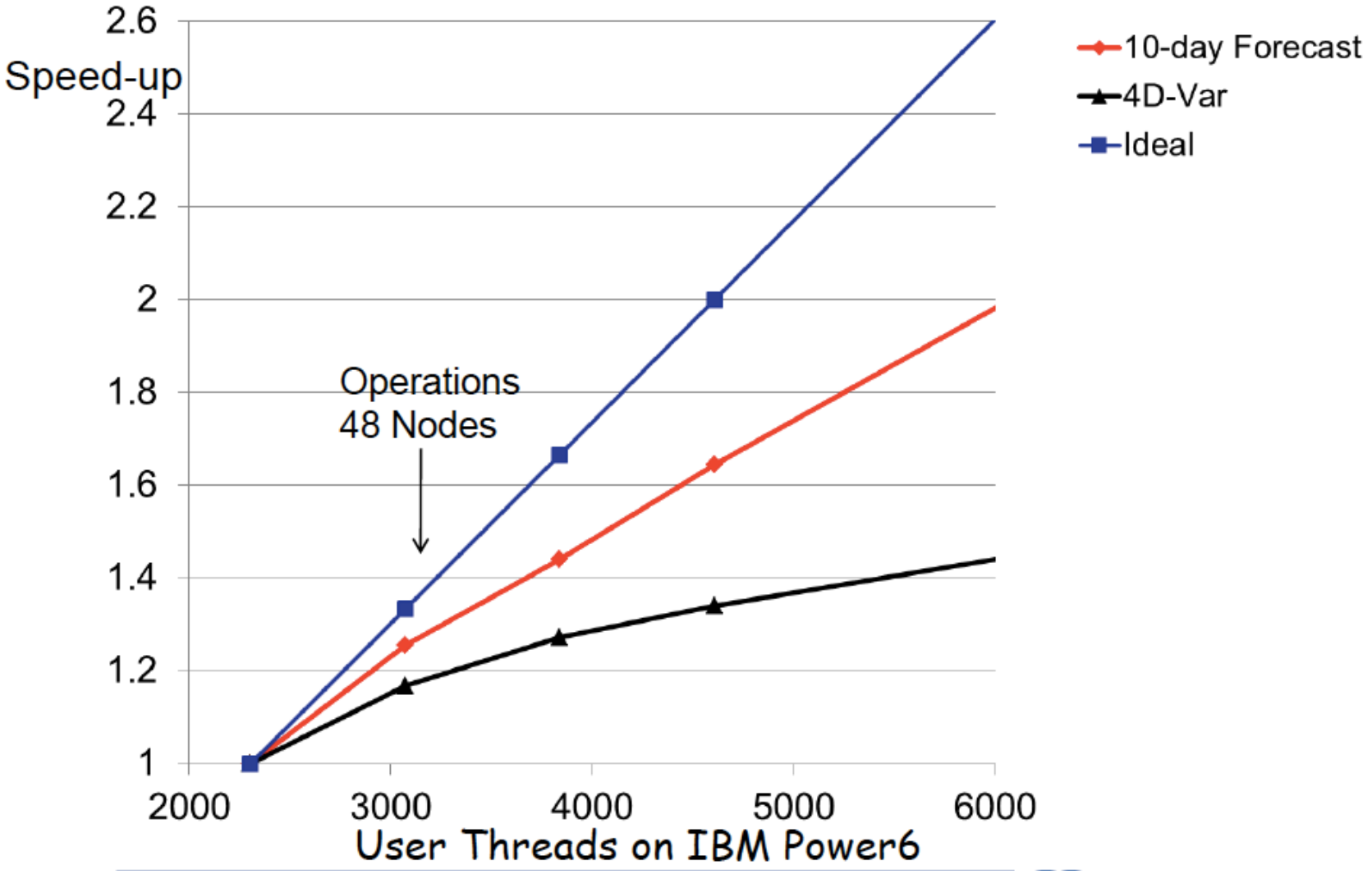


4DVAR on 48/96 nodes



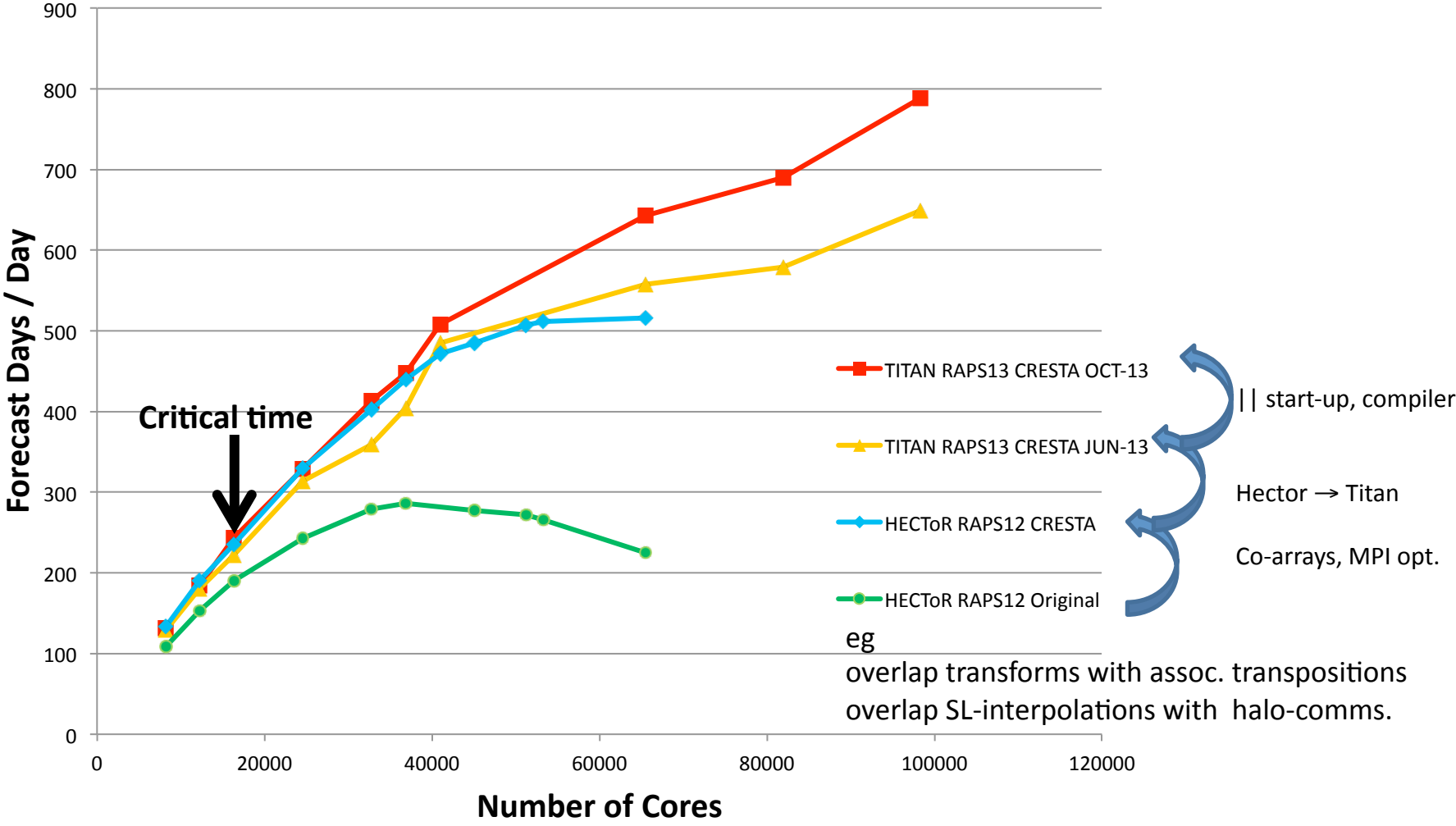
- **Analysis**
  - Sequential nature of variational data assimilation (time windows, iterations); inhomogeneous data distribution
- **Forecast**
  - Higher resolution requires smaller time steps; communication of global fields (spectral)
- **Pre-/post-processing**
  - Diversity/volume of observational data; size/speed of high resolution model output
- **Computer hardware**
  - Architecture of CPU/GPU et al./vector units; compilers; implications for code architecture

Experiments with IFS: Main components



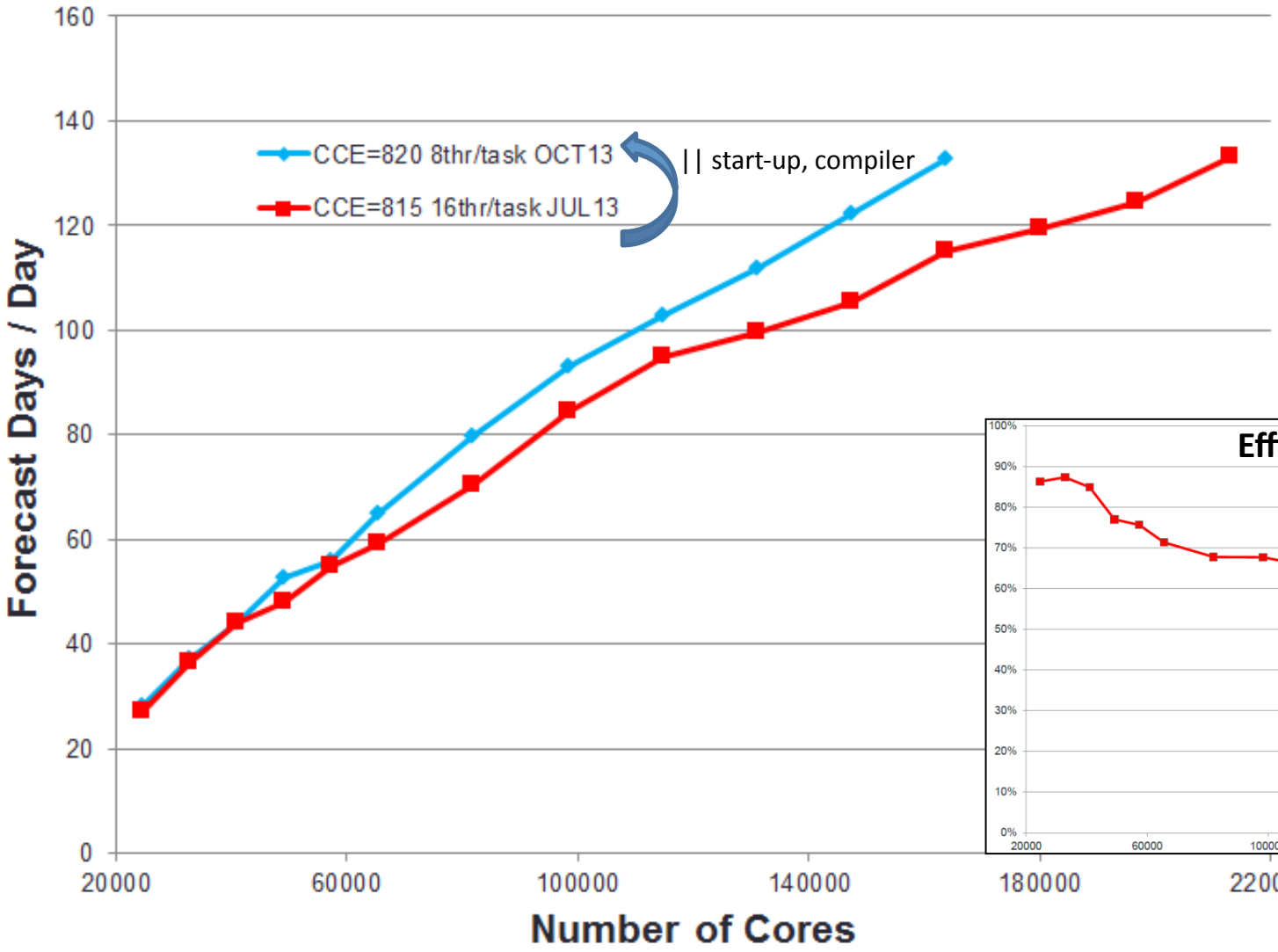
Experiments with IFS: T2047L137 (10 km)

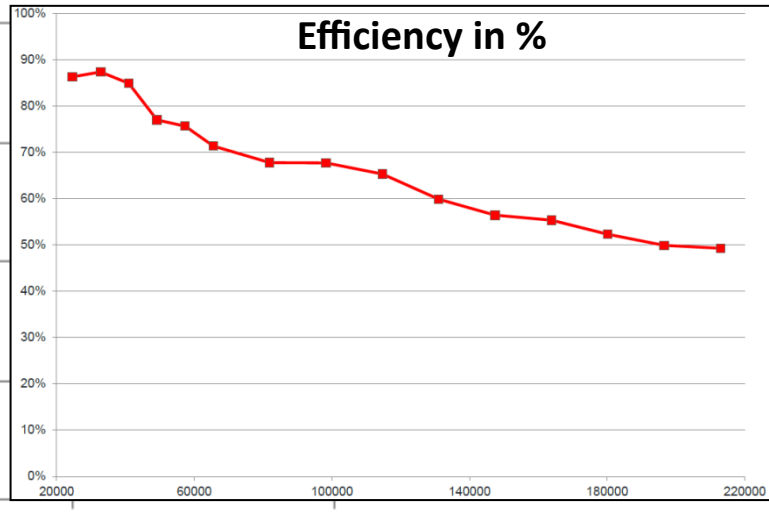
RAPS12 (CY37R3, on HECToR), RAPS13 (CY38R2, on TITAN)



Experiments with IFS: T3999L137 (5 km)

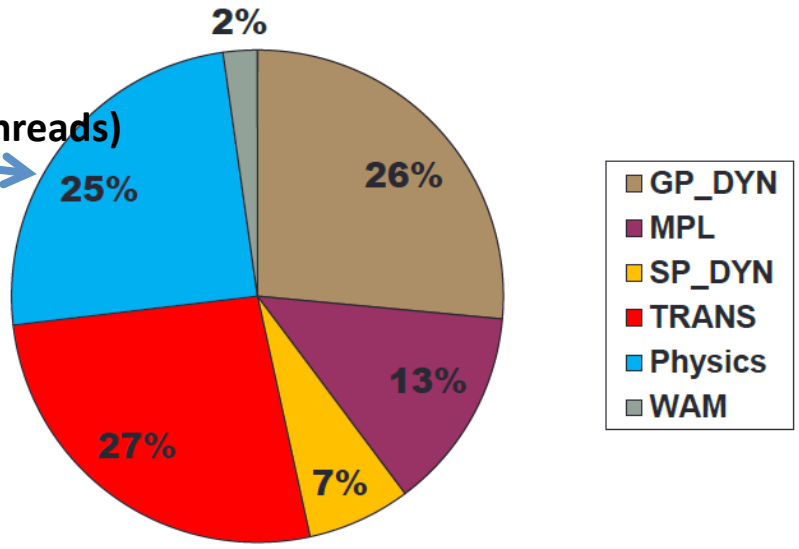
Critical time



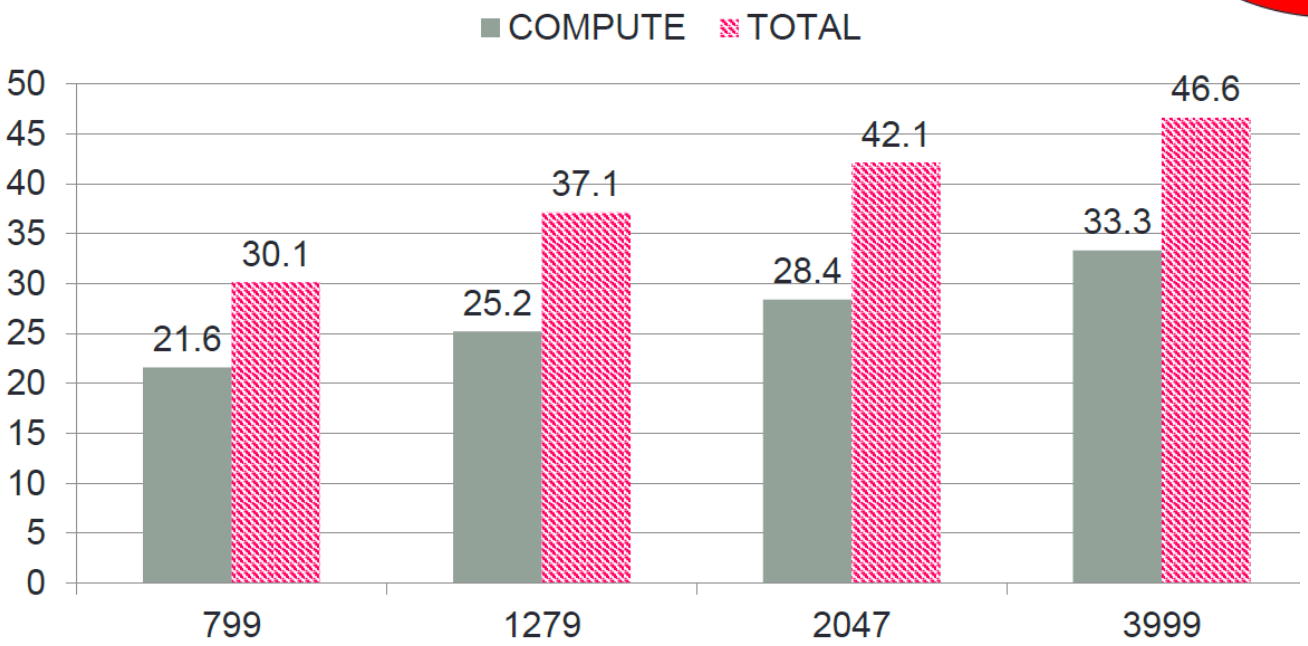


Experiments with IFS: NH Cost

T3999L91 (time step 3', 10d/4h)
(1024 MPI tasks x 16 OpenMP threads)



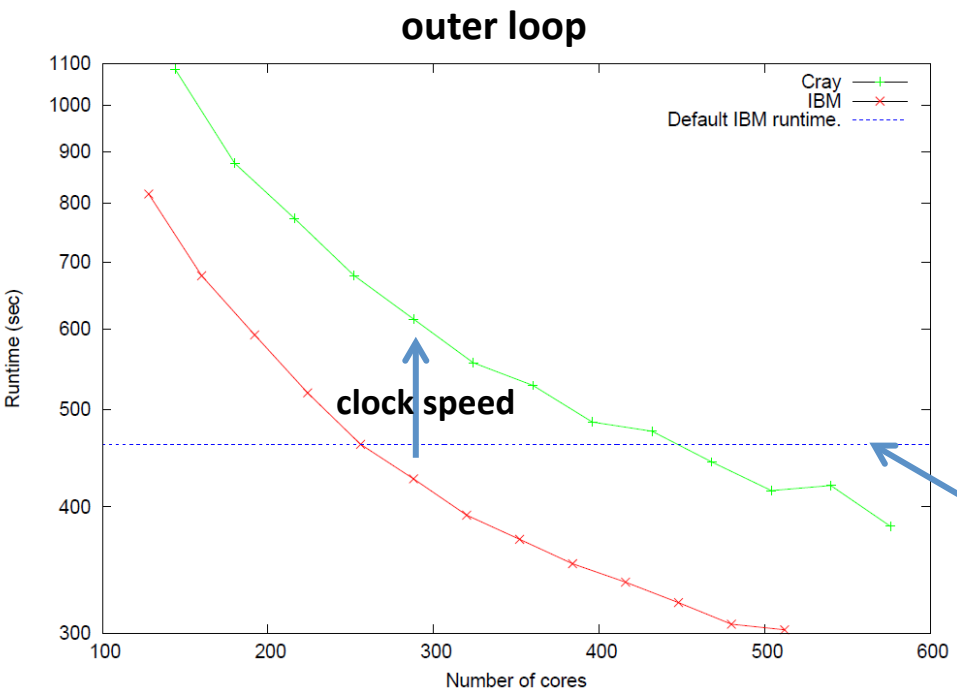
Relative cost (%) of spectral transforms
(hydrostatic = ½ non-hydrostatic)

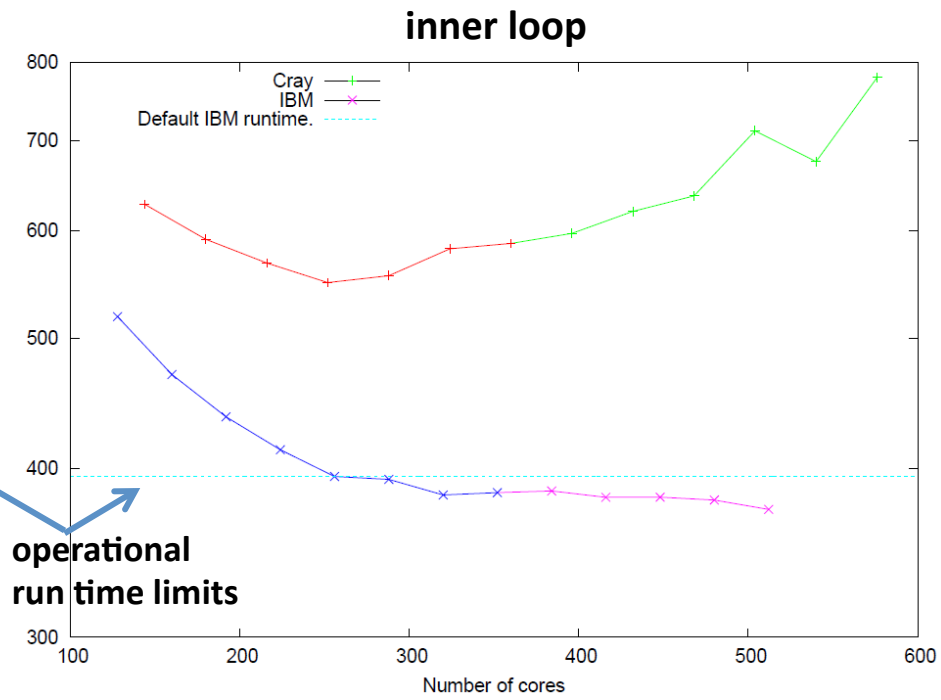


Experiments with IFS: Main components

¼ degree NEMOVAR (currently 1 degree in operations):

- Outer loop: ocean model forward integration
- Inner loop: semi-implicit scheme requires global communication in minimization, reduced by split in 2 directions





ECMWF: Scalability Programme

ECMWF scalability programme objectives:

1. to develop the future IFS combining a flexible framework for scientific choices to be made with maximum achievable parallelism,
2. to prepare for expected future technologies and their implications on code structure ensuring efficiency and code readability,
3. to develop environment/metrics for quantitative scalability assessment, through
   - coordinating ECMWF-internal resources, R&D strategy,
   - engaging with external partners (Member States, academia, HPC centres, vendors).

➔ Cray Phase-2: 2016, next HPC: 2018 (hybrid?)

What ECMWF can offer at European level:

- Experience with complex system operating under very tight time constraints
  - Operations on O(10k cores) computing & O(20 Tbyte/day) archiving
  - Continual code optimization efforts
- Scalability Programme addressing:
  - Work flows, model formulation, data assimilation algorithm, I/O
- By convention international effort

ECMWF: Scalability Programme

ECM
1.
2.
3.

be
ure
lors).

ECMWF

European Centre for Medium-Range Weather Forecasts

VACANCY NOTICE

Date of Issue: 14 March 2014

FUNCTION:	**Scientists and computer scientists to work on Scalability Programme**
	(Four positions in total in the Research and Forecast Departments)

What ECMWF can offer at European level:

- Experience with complex system operating under very tight time constraints
  - Operations on O(10k cores) computing & O(20 Tbyte/day) archiving
  - Continual code optimization efforts
- Scalability Programme addressing:
  - Work flows, model formulation, data assimilation algorithm, I/O
- By convention international effort

ECMWF: Outlook

Short-term approach:

- Overlap communication/computation: physics (esp. radiation), LT/FFT/SL-comms.
- Outsource eg LT DGEMMs, radiation to GPUs
- Extend co-array structures (portability?)
- Employ OpenMP4 once available,OpenACC
- Single executables
- Distributed I/O

→ T2047 HRES, T1023 ENS, 24h 4DVAR, Cray XC-30 phase-2

Longer-term approach:

- Alternatives to global spectral model (hybrid), local data structures
- More flexible top-level control structures (DA, i/f assimilation-model, coupling)
- Sub-windows in sequential data assimilation

→ T3999 HRES, T2047 ENS, Non-hydrostatic core, fully coupled ESM

Future consideration:

→ T7999 (2.5 km) may require 1-4 million processors to run 10d/1h

→ T3999 (5 km) 50M ensemble may also require 1-4 million processors to run 10d/1h

ECMWF: Scalability Workshop

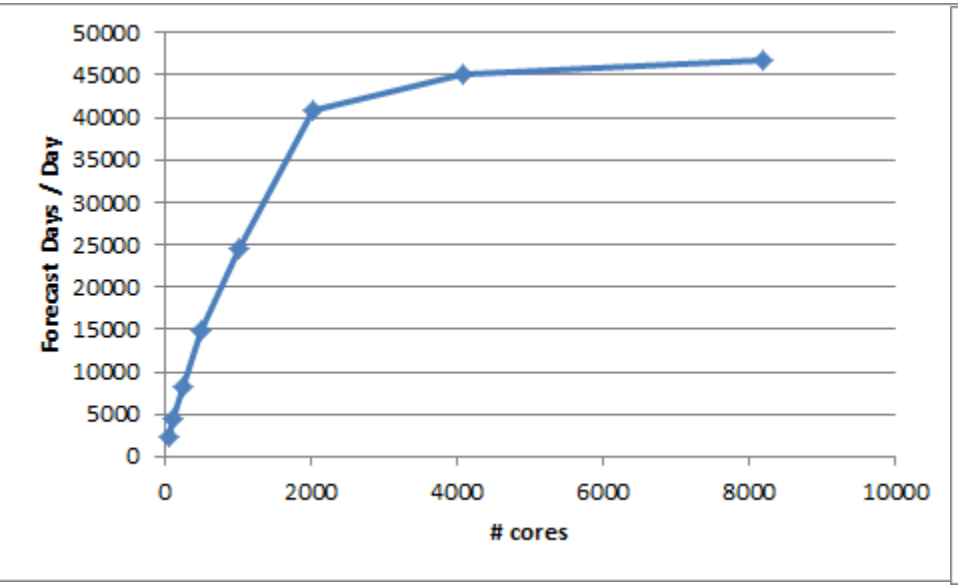
14-15 April workshop:
http://www.ecmwf.int/newsevents/meetings/workshops/2014/Scalability/

1st day: Presentations
from ECMWF, Met Office, Météo-France, DWD, Env. Canada, MeteoSwiss, Riken/AICS, LOCEAN, HIRLAM, HARMONIE, CFC, EPCC, STFC, NCAS, ENES, PRACE

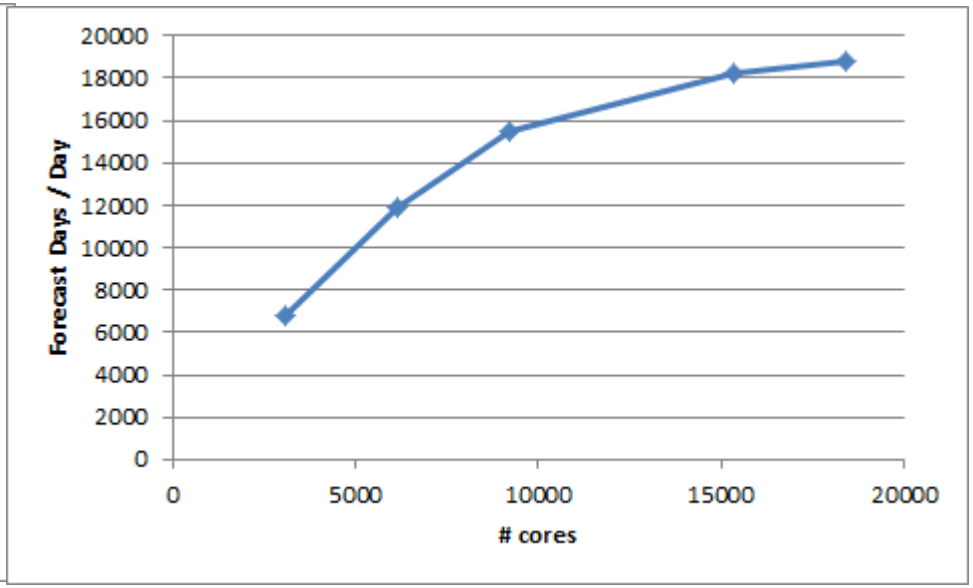
2nd day: Working groups
focussing on science & algorithms, atmosphere – composition - ocean - sea-ice coupling, data and control structures, benchmarking, hardware, languages/compilers

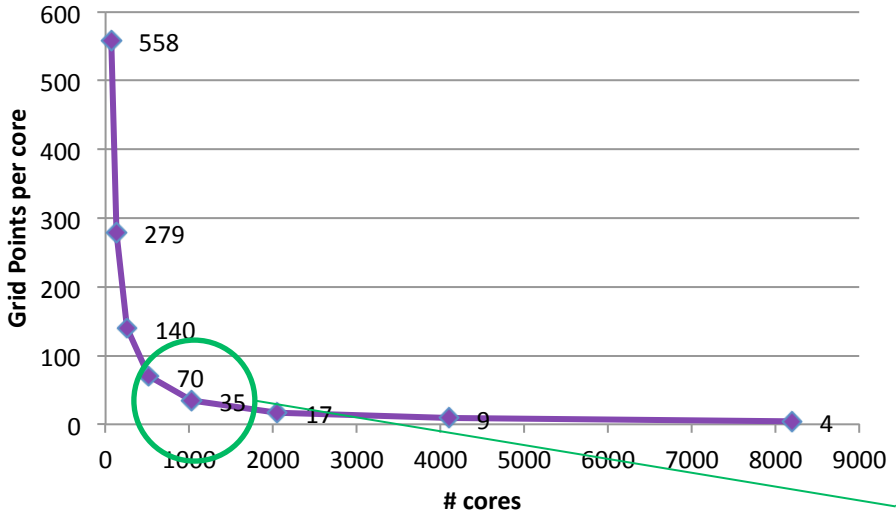
Experiments with IFS: Oakridge NRL's Titan

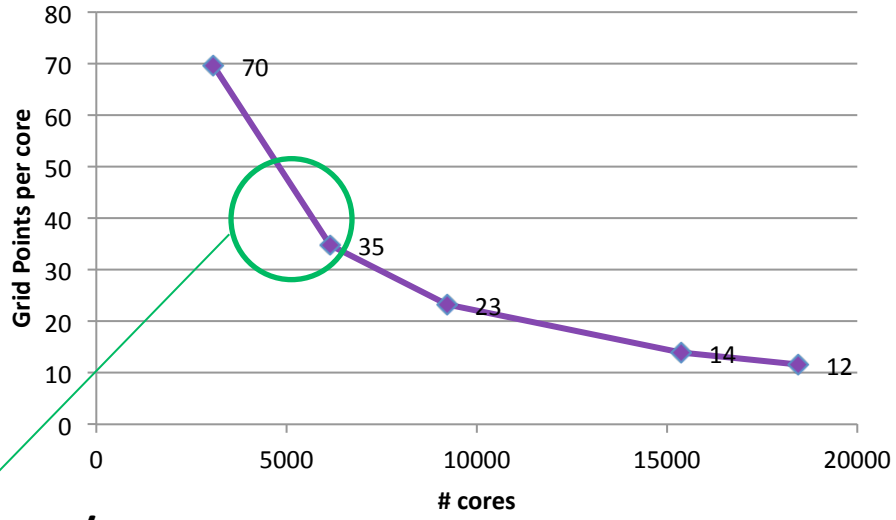
T159 (150 km) → 35k GP, 128 d/d



T399 (50 km) → 234 k GP, 51 d/d





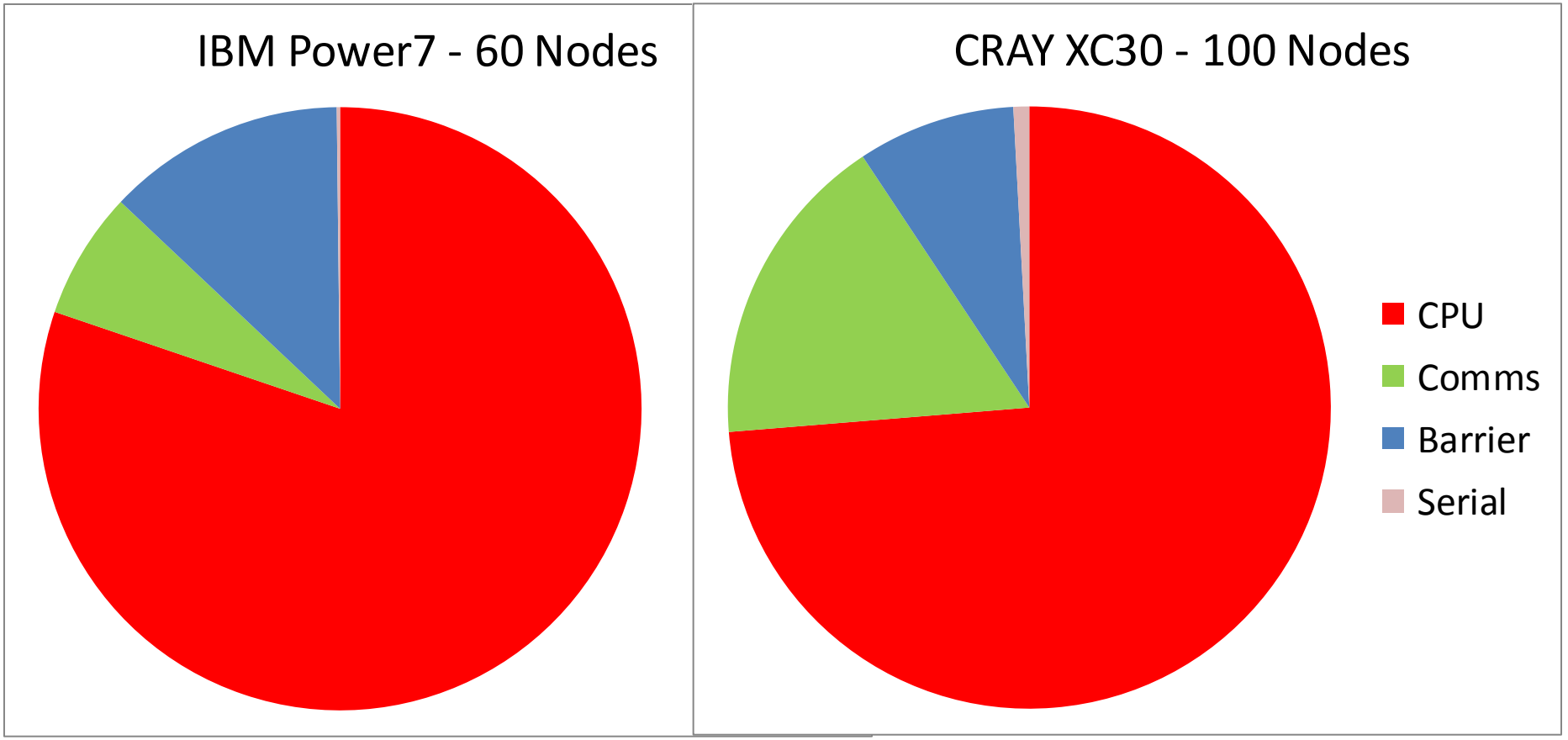


→ 50 columns/core

HPC at ECMWF

	2009-2012	2012-2014	2014-2016
	IBM Power-6	**IBM Power-7**	**Cray XC-30**
Contract phase	Phase 1	Phase 2	Phase 1
Processor	IBM Power6	IBM Power7	Intel IvyBridge
Clock	4.7 GHz	3.8 GHz	2.6 GHz
Peak Gflops /Core	18.8	30.4 (incl VSX)	20
Application nodes / cluster	262	732	~3000
Cores / cluster	8384	23424	~72000
Cores / node	32	32	24
SMT threads/core	2	2	2
Interconnect	IB – 8 links per node	IBM: HFI – 31 links per node	Cray:Aries
Parallel File-system	GPFS	GPFS	Lustre

HPC at ECMWF



2258 seconds
5.1 Tflops (8.6% peak)

2182 seconds
5.2 Tflops (10.4% peak)

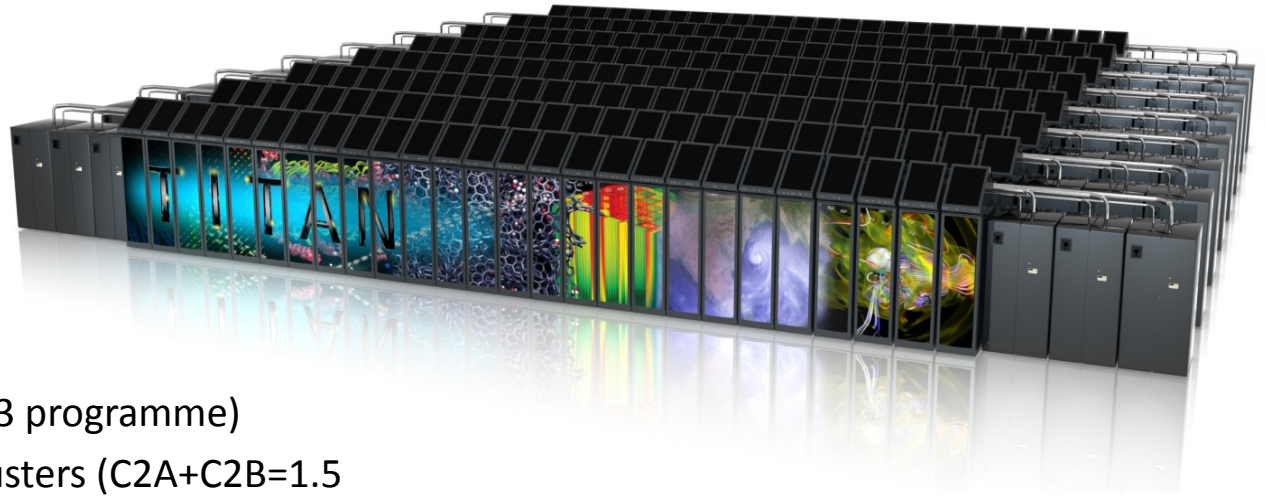
Experiments with IFS: Evolution

IFS model resolution	Envisaged Operational Implementation	Grid point spacing (km)	Time-step (seconds)	Estimated number of cores*
T1279 H	2010 (L91) 2012 (L137)	16	600	1100 1600
T2047 H	2014-2015	10	450	6K
T3999 NH	2020-2021	5	240	80K
T7999 NH	2025-2026	2.5	30-120	1-4M

***Rough estimate for the number of 'Power7' equivalent cores needed to achieve a 10 day model forecast in under 1 hour (~240 FD/D), system size would normally be 10 times this number.**

H = Hydrostatic Dynamics
NH = Non-Hydrostatic Dynamics

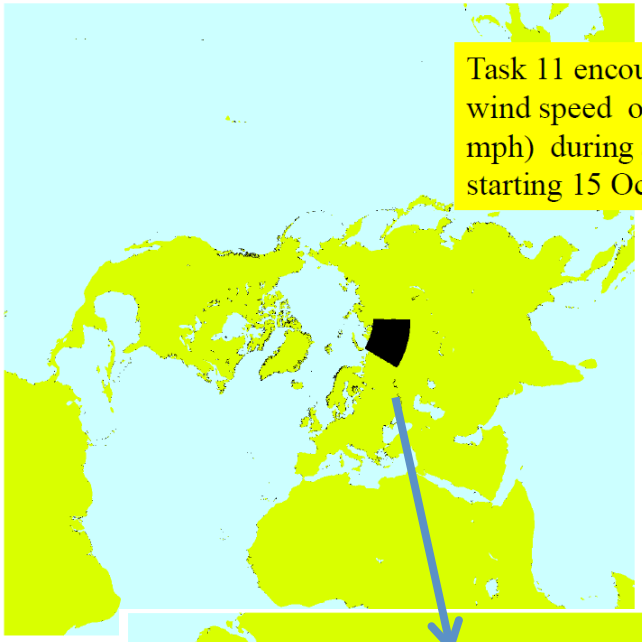
Experiments with IFS: Oakridge NRL's Titan



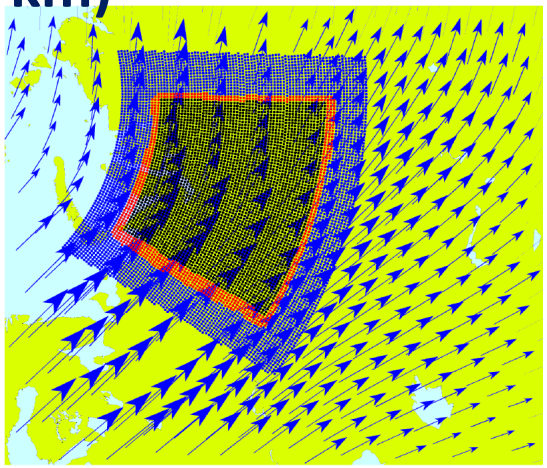
- #1 in Nov 2012 Top500 list
- CRESTA awarded access (INCITE13 programme)
- 18X peak perf. of ECMWF's P7 clusters (C2A+C2B=1.5 Petaflops)
- Upgrade of Jaguar from Cray XT5 to XK6
- Cray Linux Environment operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
- AMD Interlagos cores (16 cores per node)
- New accelerated node design using NVIDIA K20 "Kepler" multi-core accelerators
- 600 TB DDR3 mem. + 88 TB GDDR5 mem

Titan Specs	
Compute Nodes	18,688
Login & I/O Nodes	512
Memory per node	32 GB + 6 GB
# of NVIDIA K20 "Kepler" processors	14,592
Total System Memory	688 TB
Total System Peak Performance	27 Petaflops

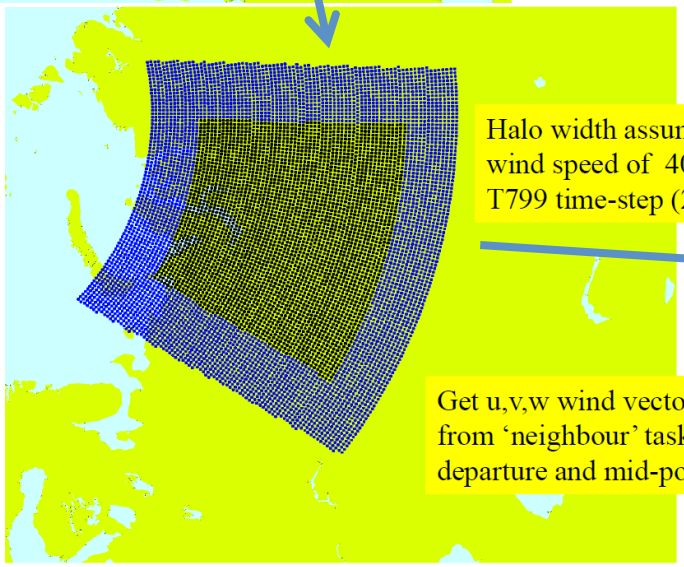
Experiments with IFS: T799L91 (25 km)



Task 11 encountered the highest wind speed of 120 m/s (268 mph) during a 10 day forecast starting 15 Oct 2004

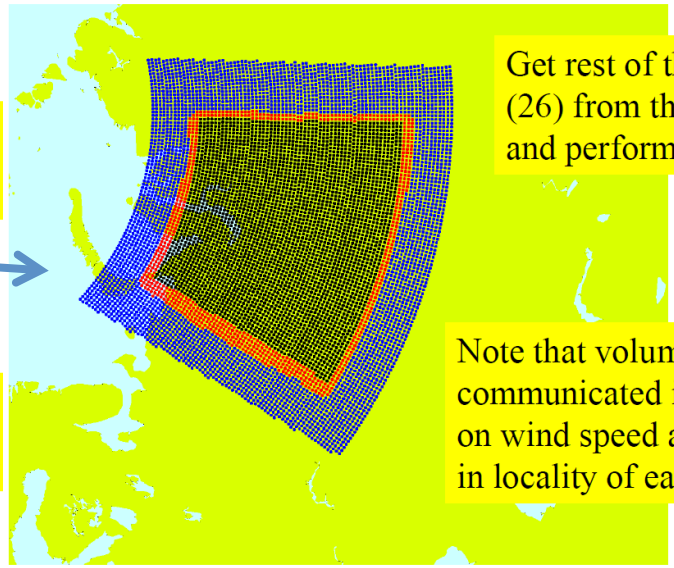


SL-halos for task 11 / 256



Halo width assumes a maximum wind speed of 400 m/s x 720 s T799 time-step (288 km)

Get u,v,w wind vector variables (3) from 'neighbour' tasks to determine departure and mid-point of trajectory



Get rest of the variables (26) from the red halo area and perform interpolations

Note that volume of halo data communicated is dependent on wind speed and direction in locality of each task