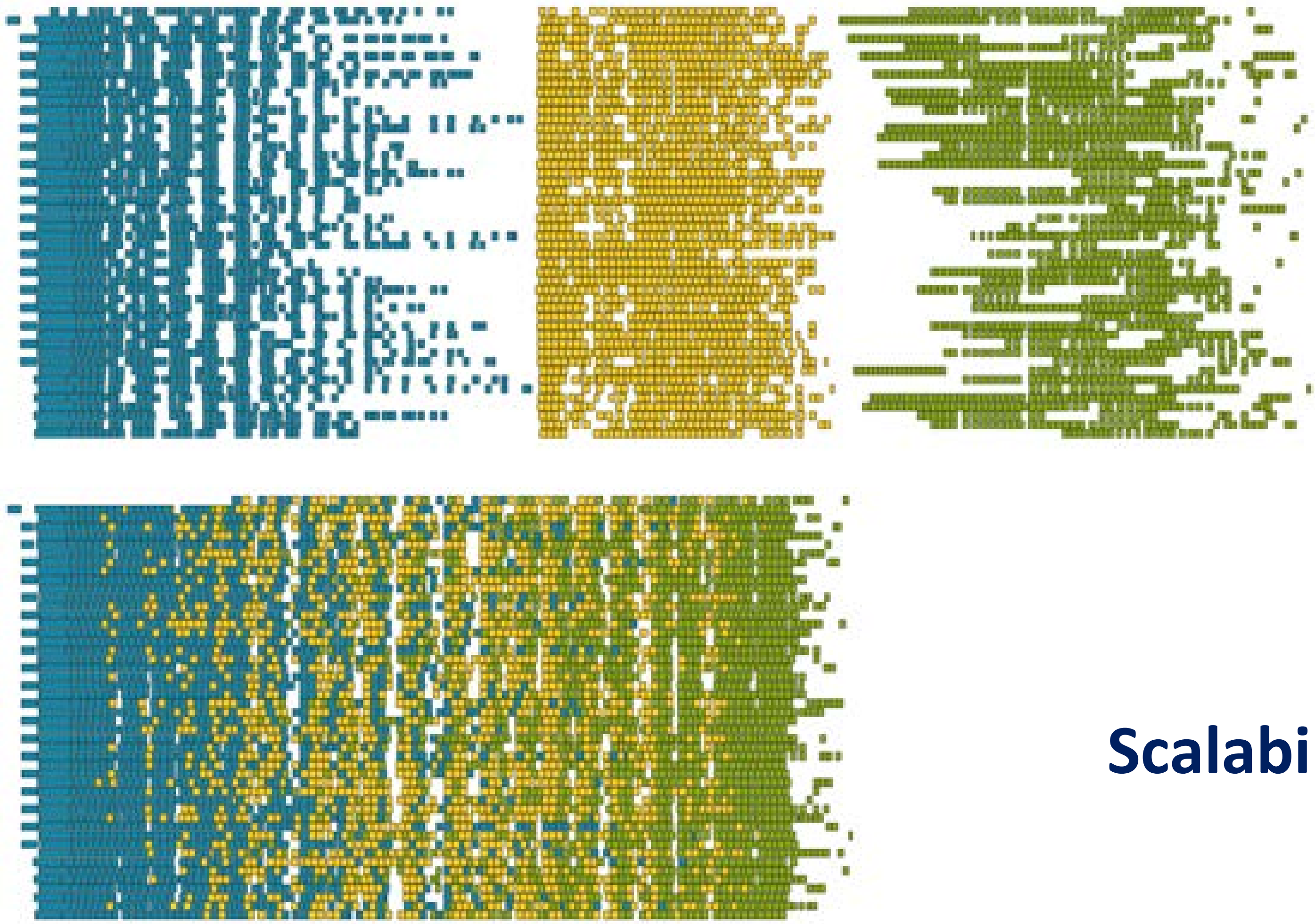


Picture: Stan Tomov, ICL, University of Tennessee, Knoxville

ECMWF
Scalability Programme

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

Next generation science developments

- Several efforts to develop the next-generation global NWP/Climate model dynamical cores (GungHo, ICON, CAM, GEM-YY, NICAM etc.):
  - unified model cores to cover (LES -) O(100m) – O(100km) range
  - with requirement for e.g. mass conservation, no computational modes, …
  - … and scalable!
- Scale-adaptive physical parameterizations:
  - grey zone (convection)
  - but also radiation (spectral-spatial-temporal, 3d effects)
  - atmospheric composition (prognostic variables)
- Coupling
  - high-resolution ocean/waves
  - sea-ice
  - land surface
- Initialization
  - sequential algorithms
  - ensemble techniques
  - and coupling

Weather and 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)

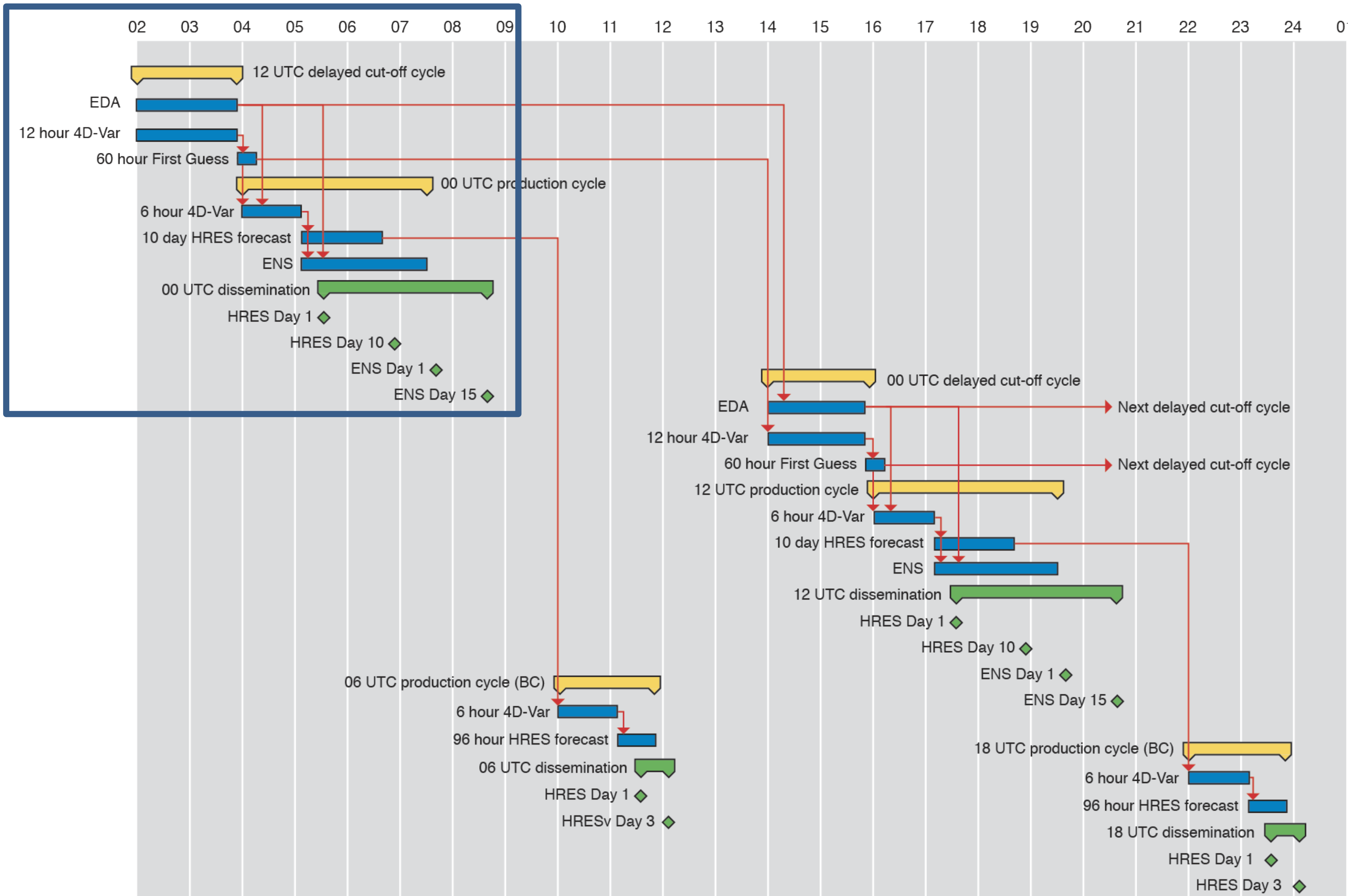
In simplified terms:

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

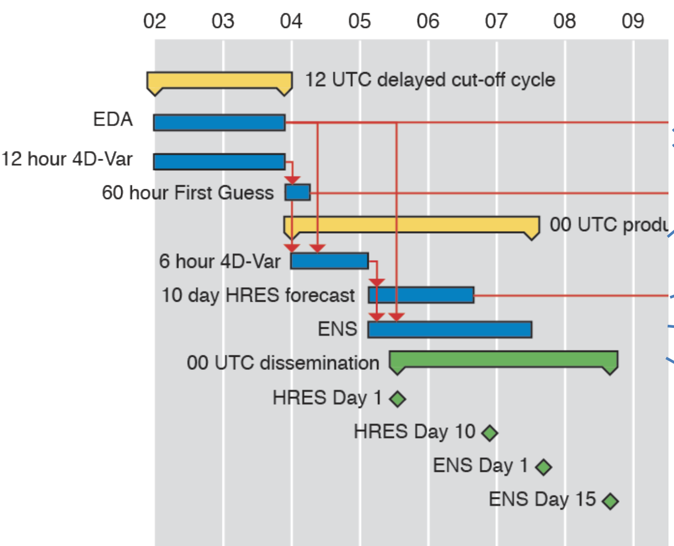
→ Earth system components etc.: weather = climate – 5-10 years

→ Main difference: long time series vs 'close-to-real-time' production

ECMWF production workflow



ECMWF production workflow



- **12h EDA**: 10 members, 2 outer loops, inner loops w/ iterations, 6h integrations, low resolution
- **6/12h 4DVAR**: 3 outer loops, inner loops w/ iterations, 6h integrations, high/low resolution, wave coupling
- Observation DB incl. feedback, ML and PL output

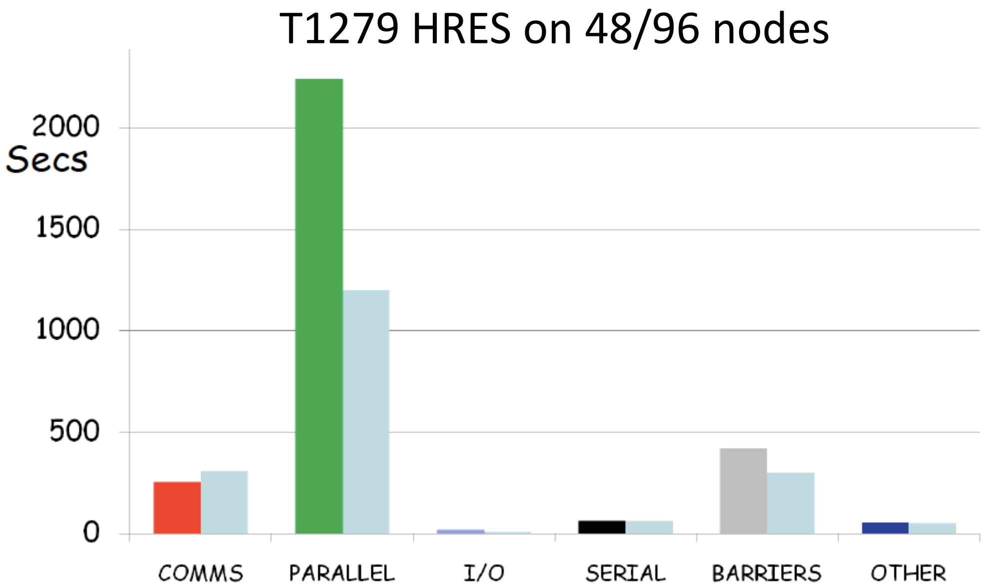
- **10d HRES**: 10d integrations, high resolution (radiation low resolution), wave coupling
- ML and PL output

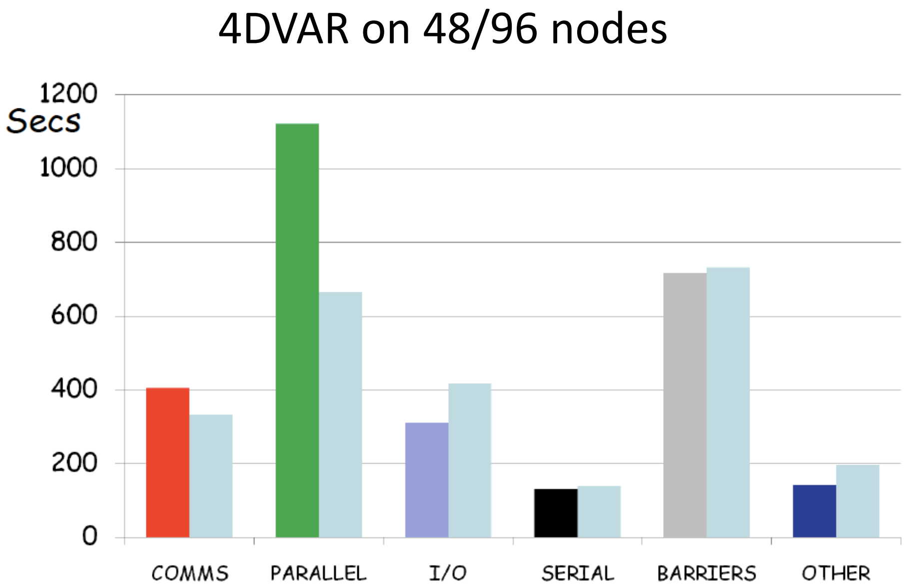
- **15/32d ENS**: 15/32d integrations, lower resolution (radiation low resolution), ocean-wave coupling,
- (2 t-steps ML and) PL output

- Archiving in MARS
- Dissemination via RMDCN

EDA	1:45h on 110 (88) nodes (IBM P7, x32 cores)
6h 4DVAR	0:40h on 120 nodes
HRES	1:15h on 60 nodes
ENS	1:55h (A) + 1:25h (B) + 0:55h (C) on 108 nodes
Product generation	max. 1:45h per suite
Dissemination	max. 1:15h per suite

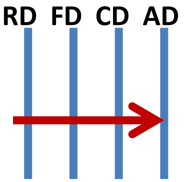
ECMWF production workflow: Main issues





- **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 (100 Gb/d); size/speed of high resolution model output (12 + 6 Tb/d)
- **Computer hardware**
  - Architecture of CPU/accelerators/vector units; compilers; implications for code design

Scalability Programme & Workshop



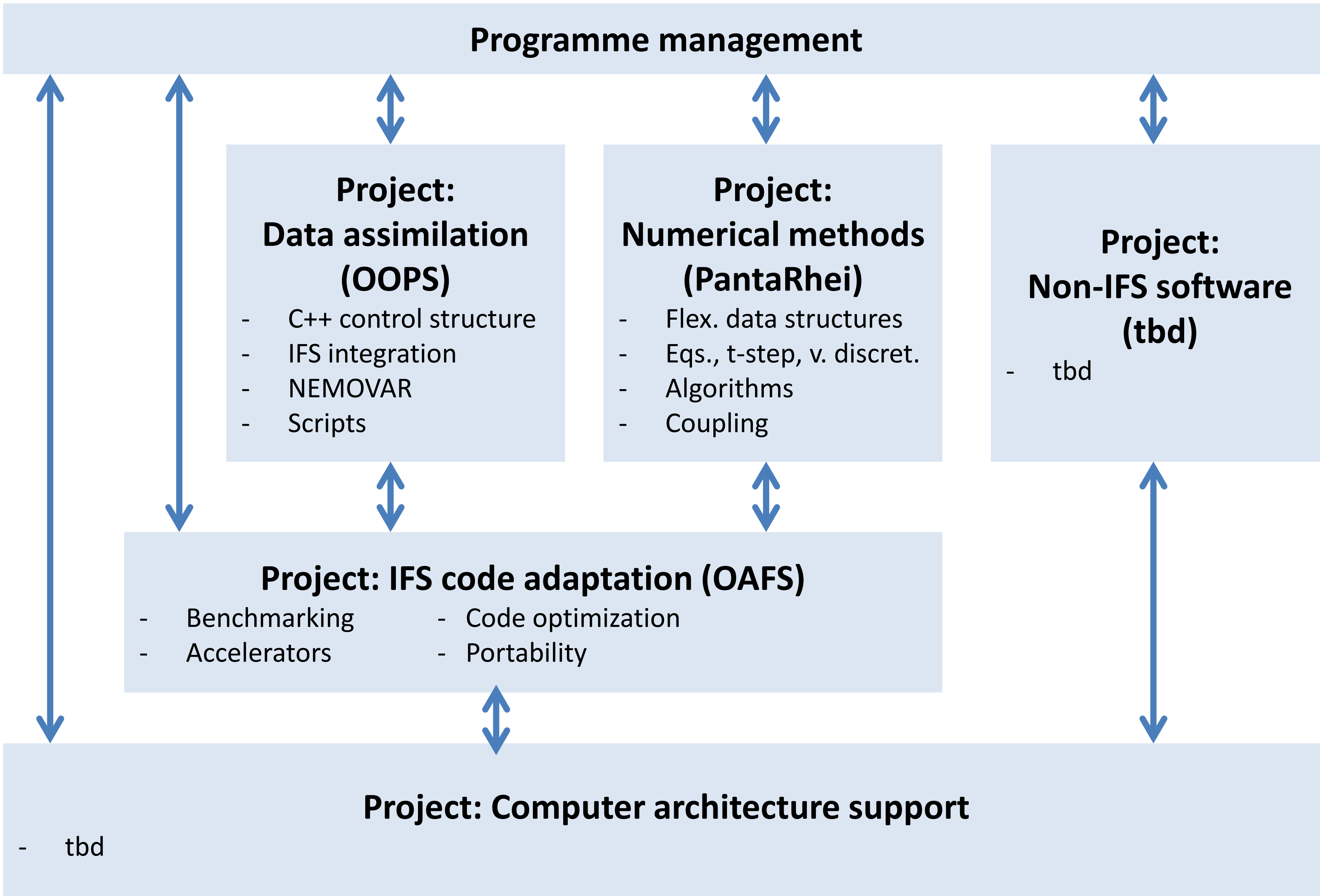
Why programme?

- Implement a formal structure at ECMWF to coordinate science & software activities across departments for efficient exa-scale computing/archiving
  - Interface with other R&D developments
  - Support future procurements
- Coordinate activities with Member States
- Coordinate activities with European HPC facilities, research centres, academia, vendors
- Coordinate with international centres

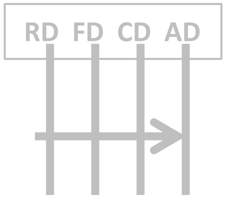
Why workshop?

- Define common areas of fundamental research towards exa-scale scalability for numerical algorithms, software infrastructure and code adaptation
- Explore the potential for common/shared code components
- Define future benchmarking strategies
- Build partnerships with science and industry
- Find opportunities for accessing external HPC resources, including novel architectures
- Explore options for consortia funded by H2020

ECMWF Scalability Programme



Scalability Programme & Workshop



Why programme?

- Implement a formal structure at ECMWF to coordinate science & software activities across departments for efficient exa-scale computing/archiving
    - Interface with other R&D developments
    - Support future procurements
- Coordinate activities with Member States
- Coordinate activities with European HPC facilities, research centres, academia, vendors
- Coordinate with international centres

Why workshop?

- Define common areas of fundamental research towards exa-scale scalability for numerical algorithms, software infrastructure and code adaptation
- Explore the potential for common/shared code components
- Define future benchmarking strategies
- Build partnerships with science and industry
- Find opportunities for accessing external HPC resources, including novel architectures
- Explore options for consortia funded by H2020

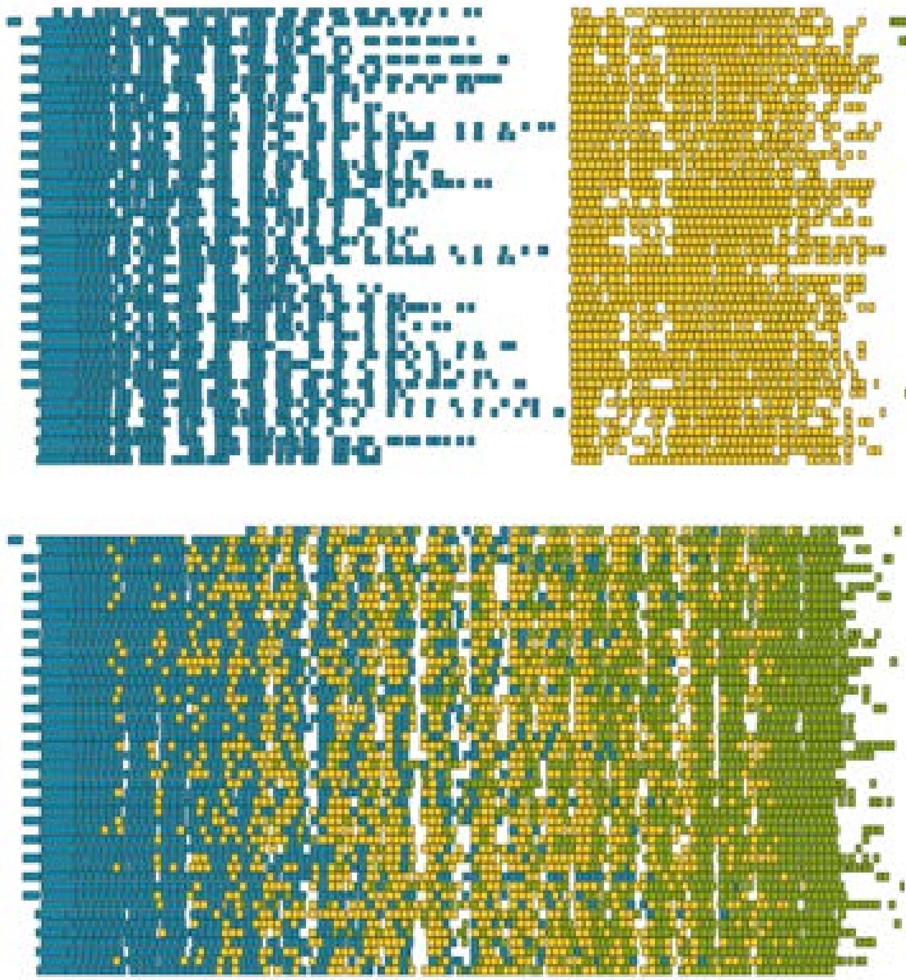
Scalability Programme & Workshop

Why programme?

- Implement a formal structure at ECMWF to coordinate science & software activities across departments for efficient exa-scale computing/archiving
  - Interface with other R&D developments
  - Support future procurements
- Coordinate activities with Member States
- Coordinate activities with European HPC facilities, research centres, academia, vendors
- Coordinate with international centres

Why workshop?

- Define common areas of fundamental research towards exa-scale scalability for numerical algorithms, software infrastructure and code adaptation
- Explore the potential for common/shared code components
- Define future benchmarking strategies
- Build partnerships with science and industry
- Find opportunities for accessing external HPC resources, including novel architectures
- Explore options for consortia funded by H2020



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

Working Groups

Peter Bauer

Scalability Project: Workshop

- 3 working groups of about 20 participants, chaired / assisted by:
  1. John Michalakes (NCEP) / Nils Wedi (ECMWF) → Meeting room 1
  2. Alain Joly (Météo-France) / Deborah Salmond (ECMWF) → Mezzanine room
  3. Paul Selwood (Met Office) / Mike Hawkins (ECMWF) → Council chamber

- Each working group will deal with the same set of questions on topics: general, workflow, scientific flexibility/choices, numerical techniques/libraries, hardware/compilers, I/O, benchmarking

- Questions can be selected, discarded, changed, new questions can be added

- Working groups have about 4 ½ hours:
  - Recommendations for above topics
  - Recommendations for (i) joint community efforts, (ii) ECMWF focus

- Plenary will be in Lecture Theatre on Tuesday at 16:00, 15' presentation per working group plus discussion

- Post-workshop: Report, Ingestion in ECMWF Programme definition, Common projects

Scalability Project: Workshop

General:

- Is the opportunity of Exa-scale computing power fundamentally changing the way we do NWP forecast and analysis? If yes, at what anticipated time-scale do we expect the change?
- What are common components of the NWP system that may be shared between ECMWF, other centres, the climate community, regional applications?
- What should be the European approach to strengthen industry – HPC centre – science – application chain?
- Which partnerships will optimize funding opportunities in Horizon 2020?

Workflows:

- E.g. what are the main bottlenecks in current workflows?
  - Climate: long time series,
  - NWP: critical-path production and dissemination schedules,
  - Single centre vs distributed approach (few centralized HPC/archiving facilities, many users)

Scalability Project: Workshop

Scientific flexibility/choices:

E.g. which are the priorities between complexity, resolution, ensembles given scalability limitations?

Numerical techniques/libraries:

E.g. what is the trade-off between accuracy and energy efficiency (e.g. double vs single precision)?

Hardware/compilers:

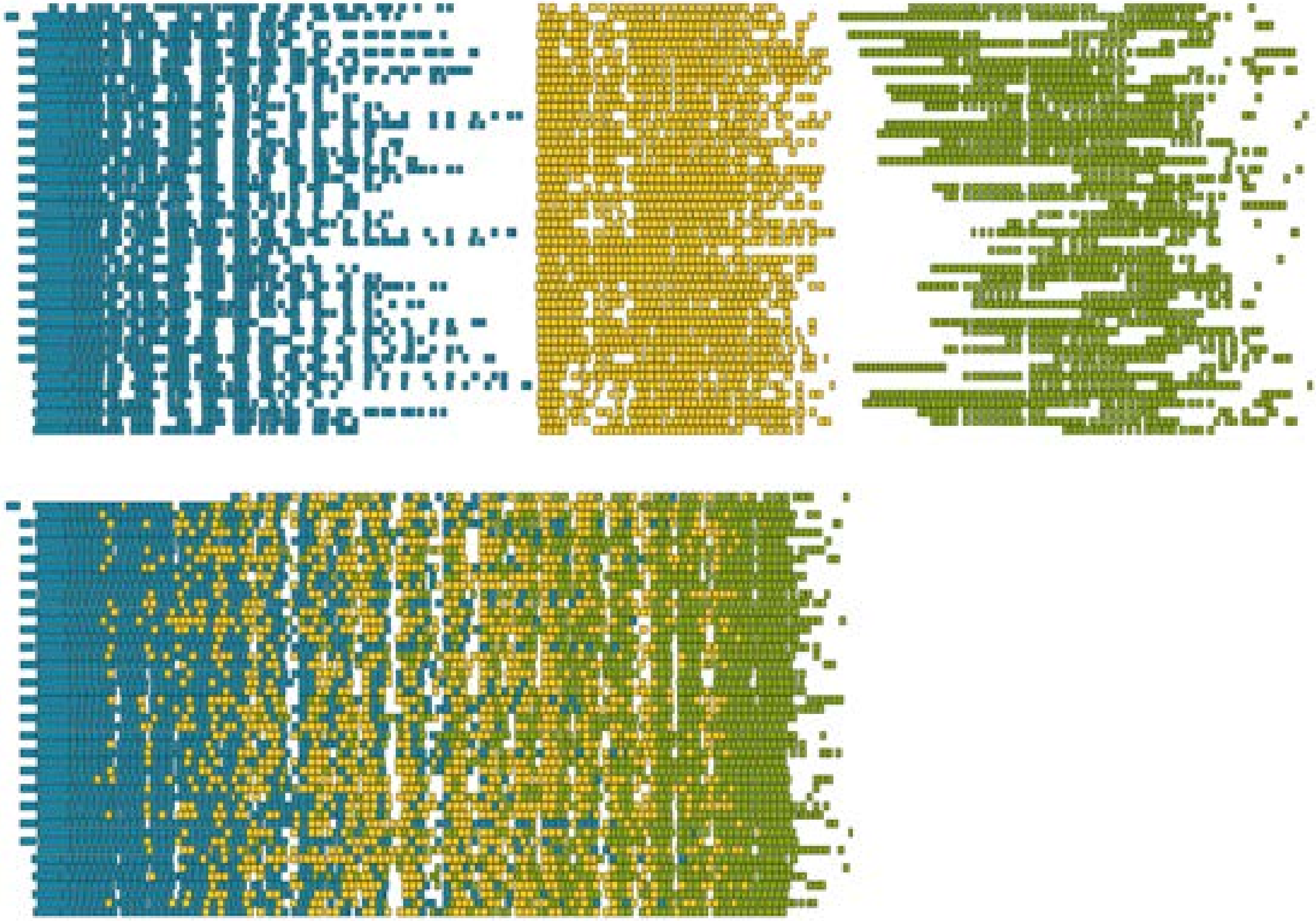
E.g. how will the other components of an exa-scale system cope, e.g. Operating System, resource scheduler, workflow management, file system?

I/O:

E.g. what needs to be archived/disseminated, what can be post-processed on the fly or recalculated?

Benchmarking:

E.g. which components of the workflow should be benchmarked separately and how?



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

Back-up

Peter Bauer

ECMWF data processing

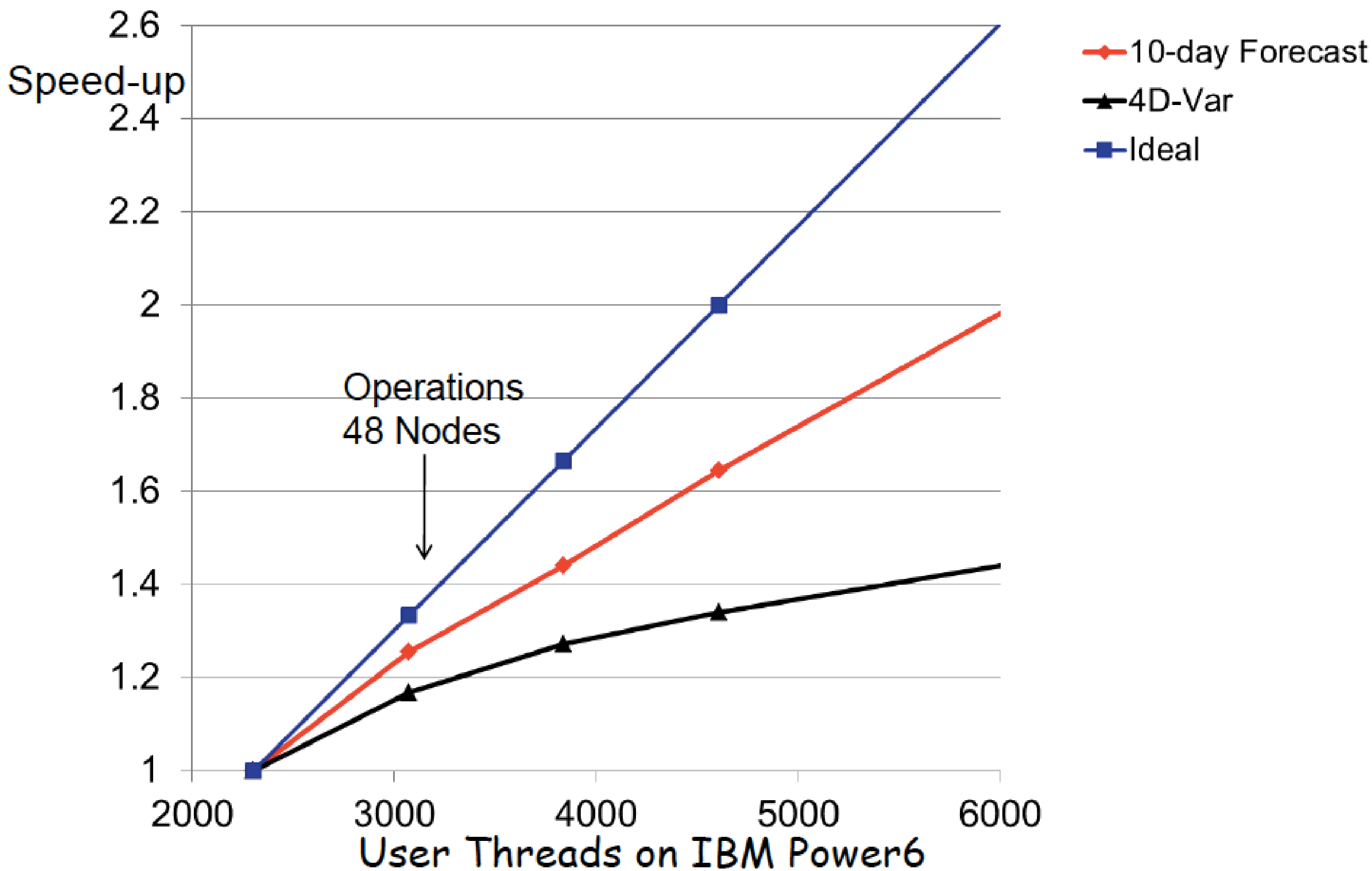
- **Observations per day: 100 Gbyte**
  Observations need to be re-processed in 30 minutes should the database be lost. In a regular situation ca. 30-50 GByte need to be transferred and pre-processed in less than 20 minutes. Feedback slightly larger but no time constraint.

- **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).

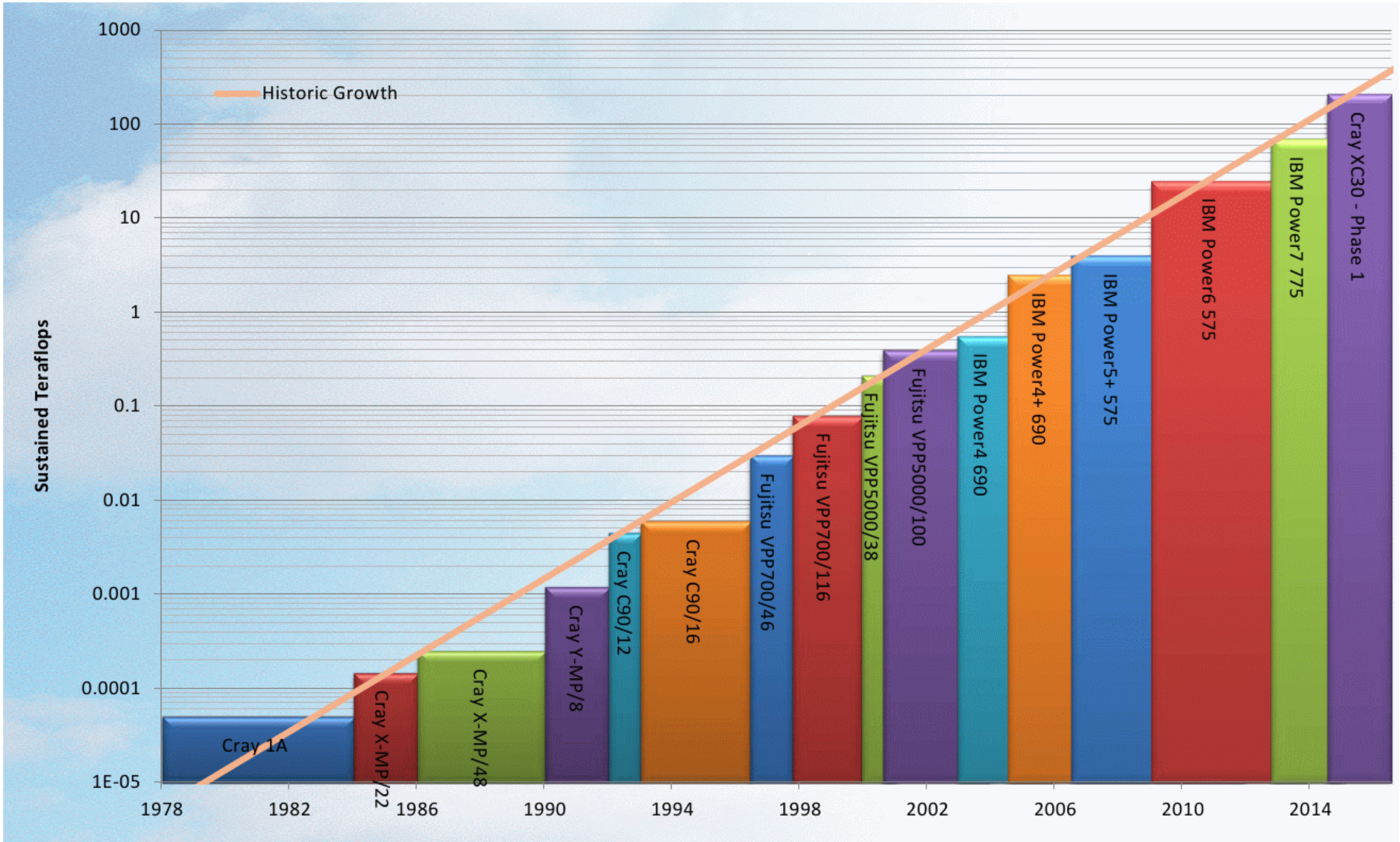
- **Products generated per day: 6 Tbyte**
  Product generation needs to run alongside the model, i.e. in addition to writing the above model output. Production generation reads the data, processes it and writes it out again. The total elapsed time is identical, 3.5 (2.5) hours.

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

Experiments with IFS: Main components



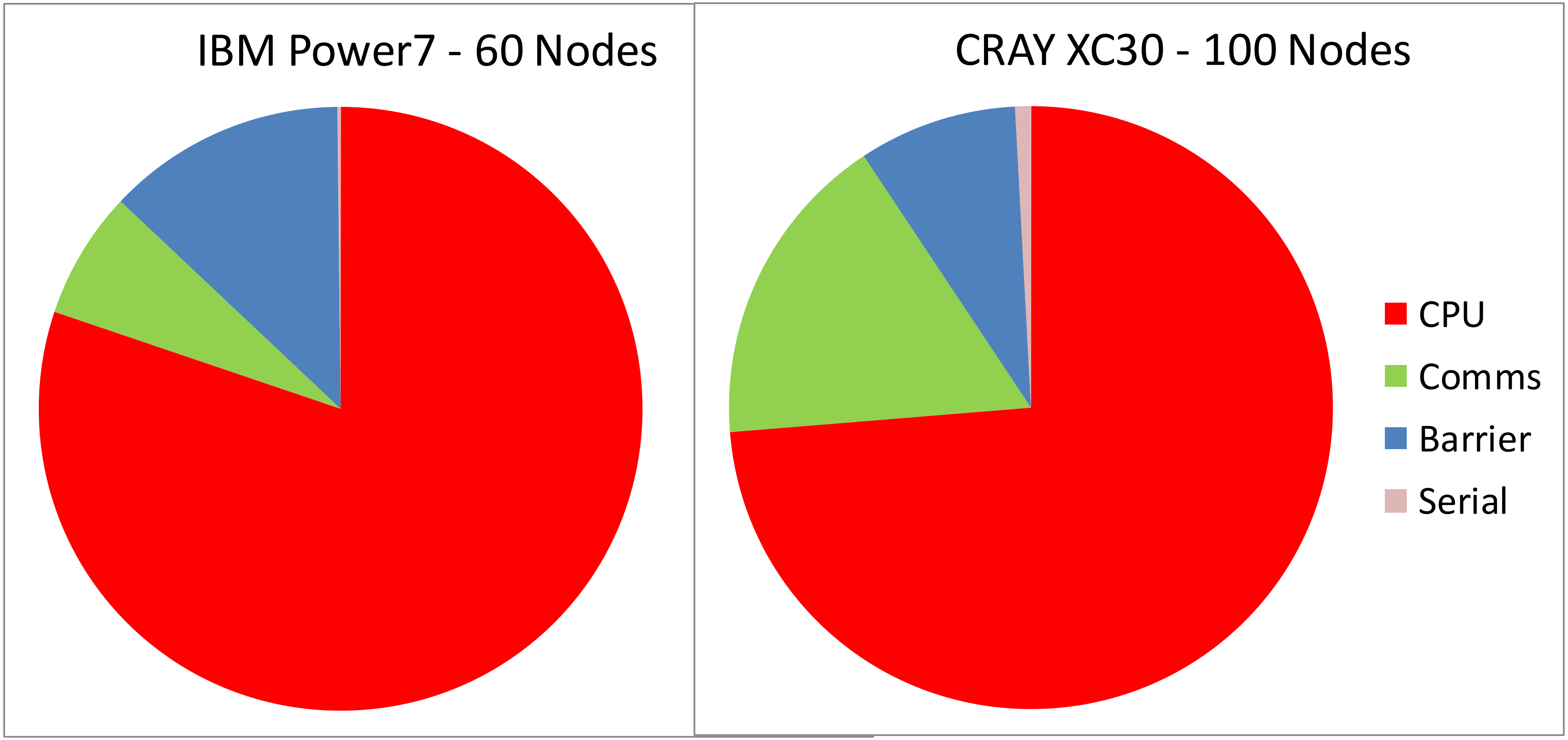
ECMWF HPC history



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

HPC at ECMWF



2258 seconds
5.1 Tflops (8.6% peak)

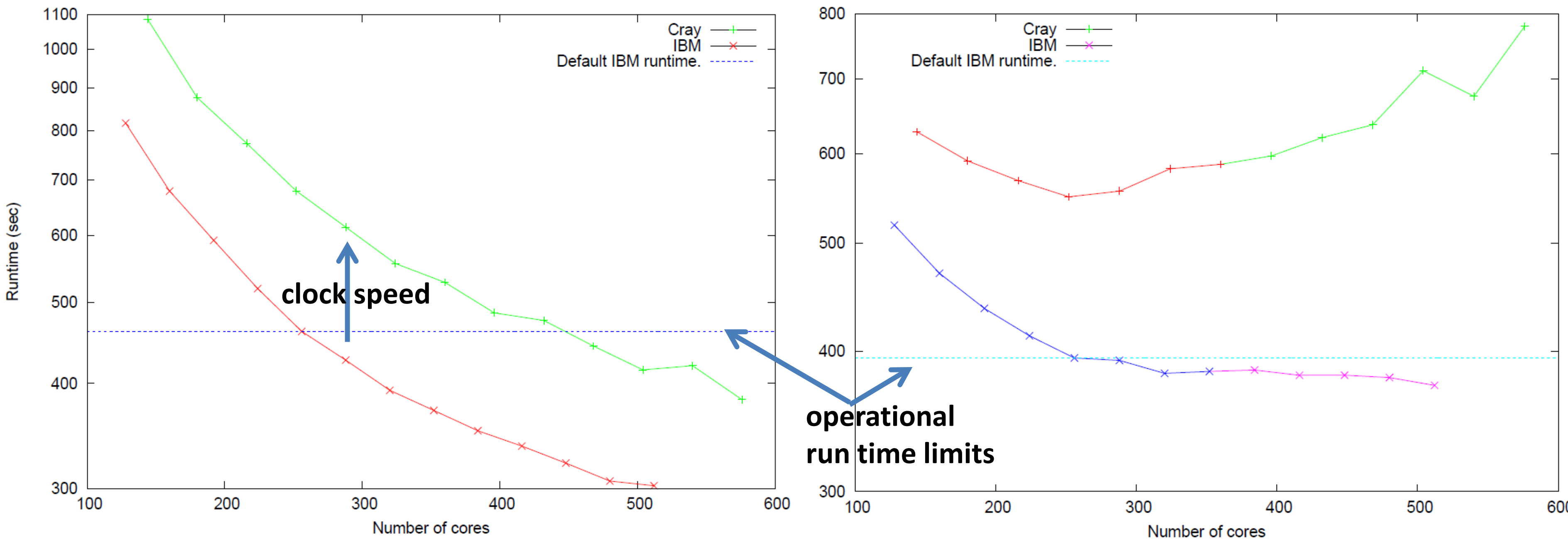
2182 seconds
5.2 Tflops (10.4% peak)

Experiments with IFS: Main components

¼ degree NEMOVAR (currently 1 degree in operations):

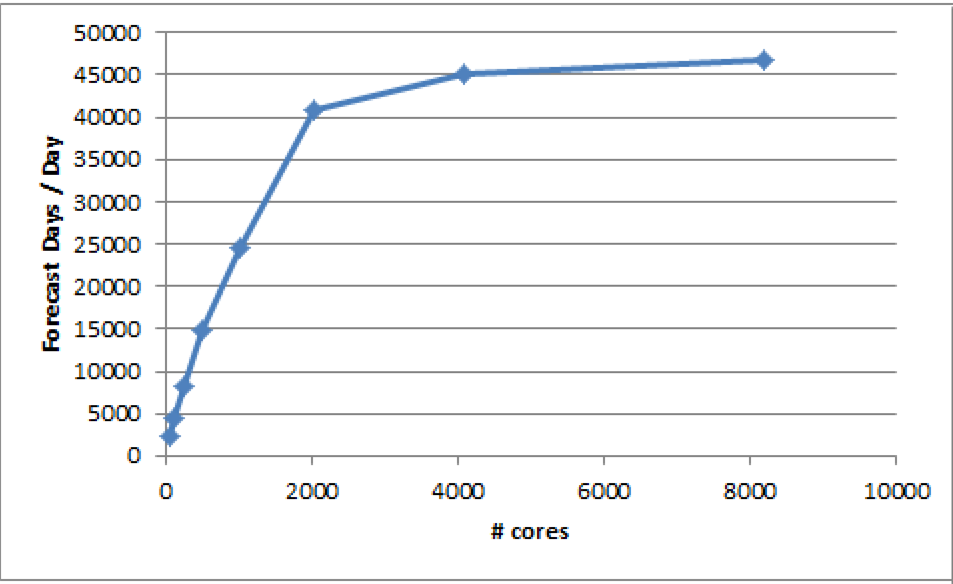
- Outer loop: ocean model forward integration
- Inner loop: semi-implicit scheme requires global communication in minimization, reduced by split in 2 directions

outer loop **inner loop**

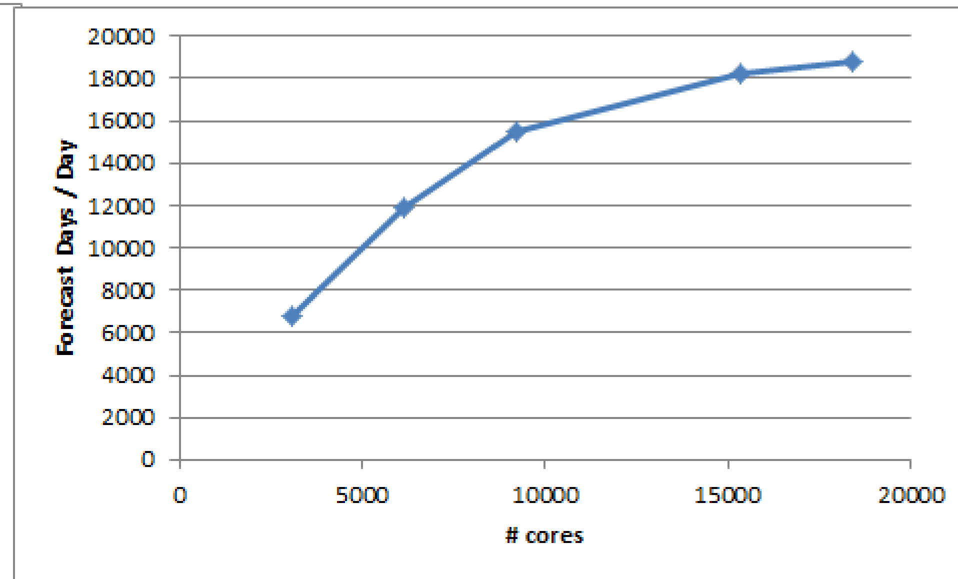


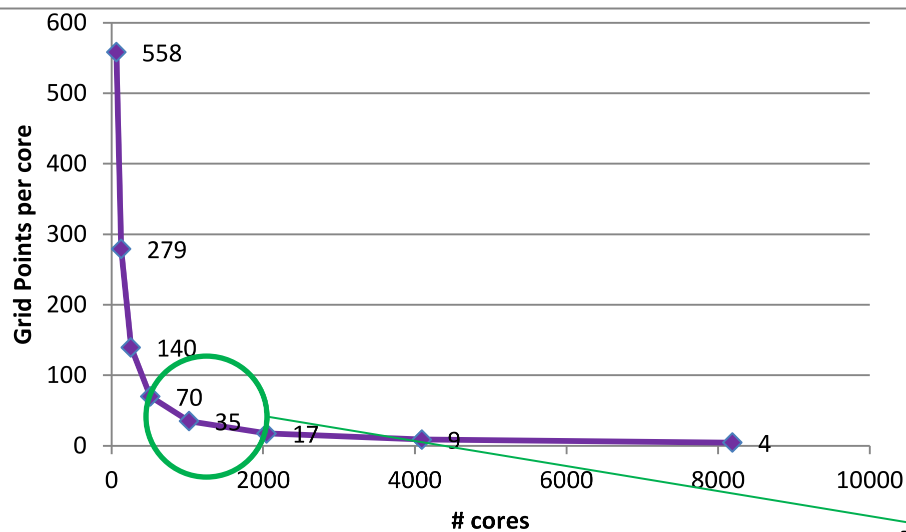
Experiments with IFS: Oakridge NRL's Titan

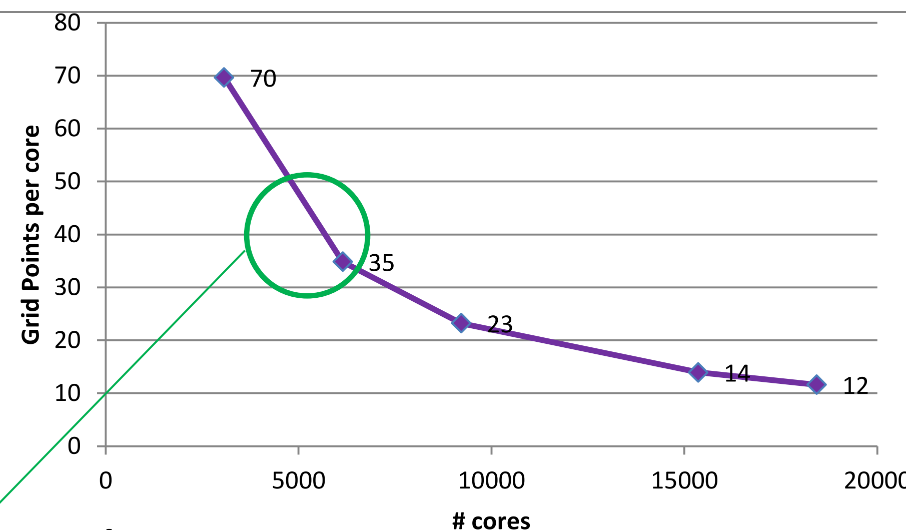
T159 (150 km) → 35k GP, 128 d/d



T399 (50 km) → 234 k GP, 51 d/d



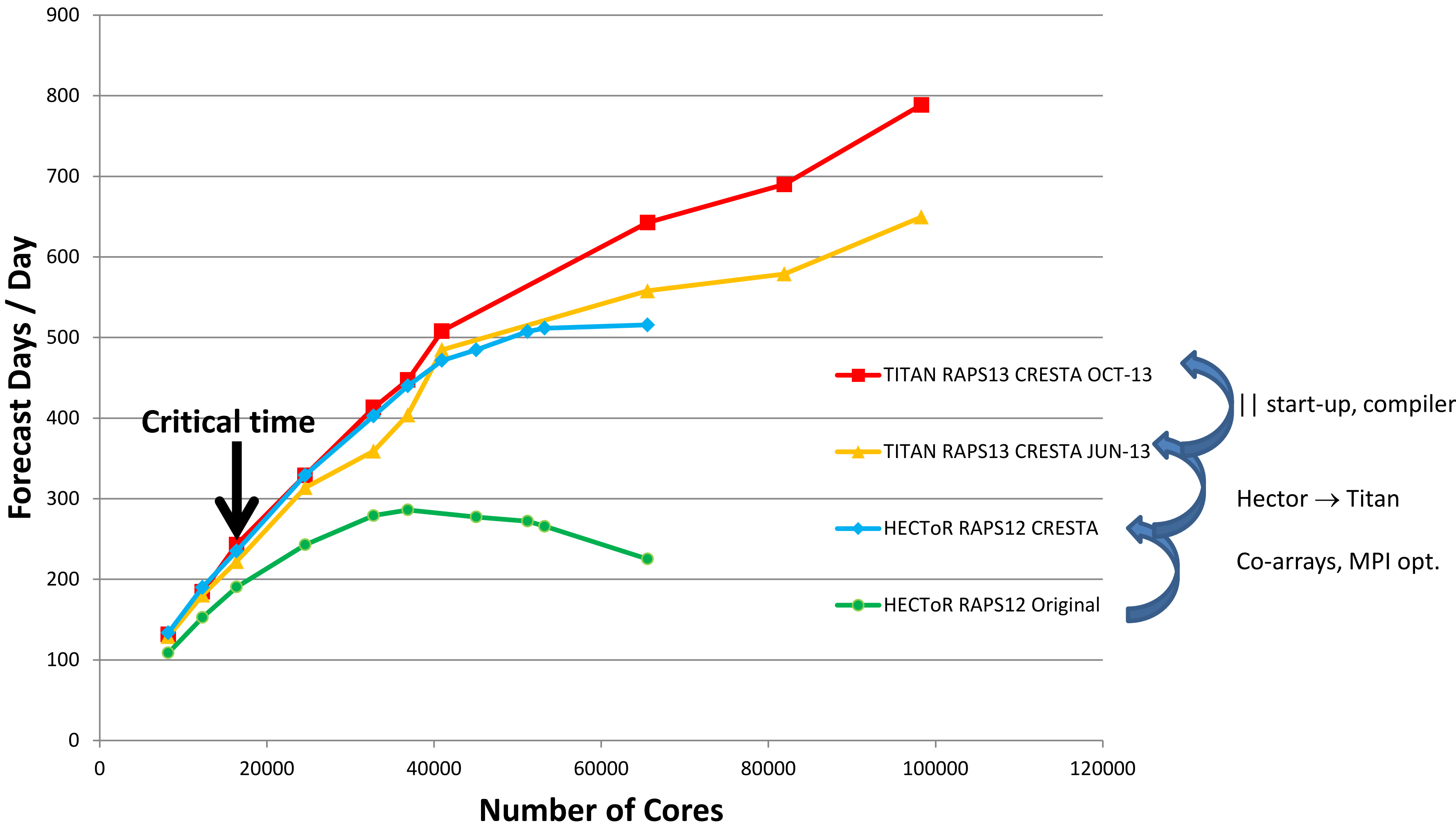




→ 50 columns/core

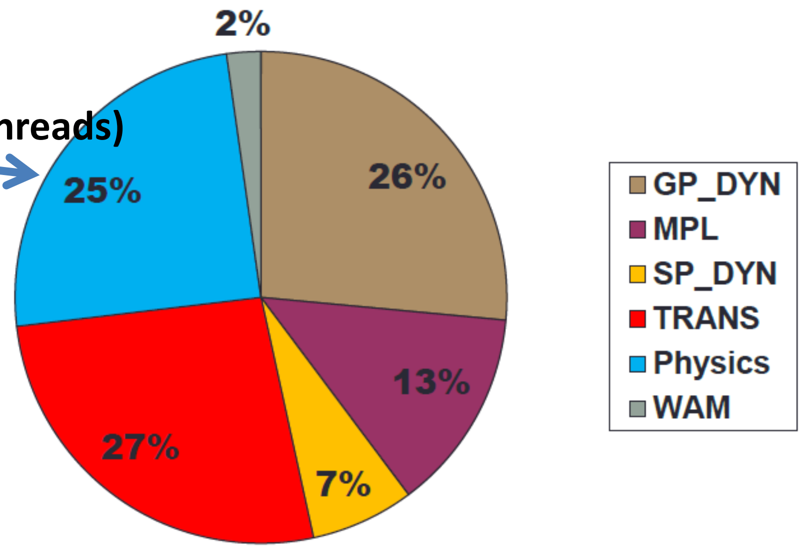
Experiments with IFS: T2047L137 (10 km)

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



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)