
MISTRAL
the new DKRZ system
ESiWACE
a new „Center of Excellence"
Joachim Biercamp
DKRZ

MISTRAL

ESiWACE

DKRZ

MISTRAL

(The new DKRZ super computer system)

1985: Control Data Cyber-205

- 1 processor
- 0.2 Gigaflops
- 0.03 Gigabyte memory

1988: Cray 2S

- 4 processors
- 2 Gigaflops
- 1 Gigabyte memory

1994: Cray C-916 „Sea"

- 16 processors
- 16 Gigaflops
- 2 Gigabyte memory
- 128 Gigabyte disc space
- 10 Terabyte tape archive

1985 1988 1994 2000

2002: NEC SX-6 „Hurrikan"

- 192 processors
- 1.5 Teraflops
- 1.5 Terabyte memory
- 60 Terabyte disc space
- 3.4 Petabyte tape archive

2009: IBM Power6 „Blizzard"

- 8500 processors
- 158 Teraflops
- 20 Terabyte memory
- 6 Petabyte disc space
- 60 Petabyte tape archive

2015/16: bullx B700 DLC „Mistral"

- 36000 - ca 80000 processors
- 1.5 - 3+ Petaflops
- 120 - 240 Terabyte memory
- 20 - 50 Petabyte disc space
- up to 500 Petabyte tape archive

2002 2009 2015 2016

3D-atmosphere or ocean only models
First coupled models (Atm. + mixed layer ocean)
~ 5 deg grid spacing
Simulation time: month to a few
4 gigabyte of data

1994: Cray C-916
„Sea"

- 16 processors
- 16 Gigaflops
- 2 Gigabyte memory
- 128 Gigabyte disc space
- 10 Terabyte tape archive

1985 1988 1994 2000

2002: NEC SX-6
„Hurrikan"

- 192 processors
- 1.5 Teraflops
- 1.5 Terabyte memory
- 60 Terabyte disc space
- 3.4 Petabyte tape archive

2009: IBM Power6

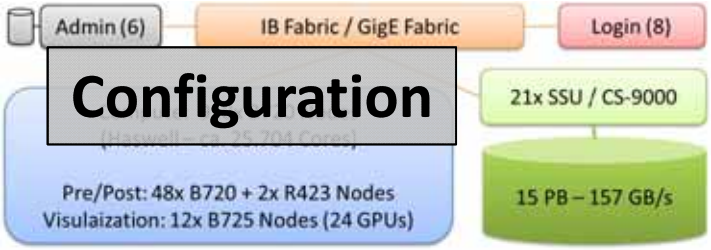
2015/16: bullx B7[illegible]DLC

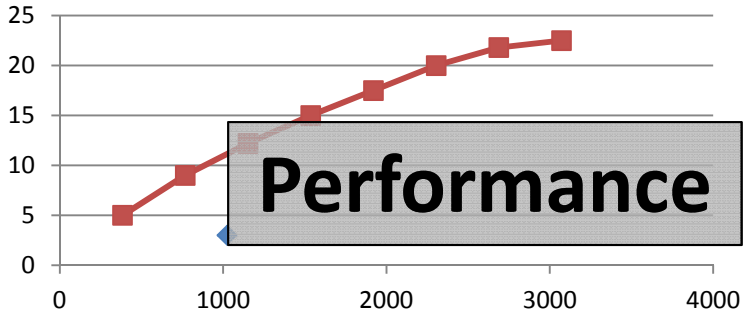
Coupled ESM ~ .5 deg resolution
Simulation time: 100-100 years
Several petabyte of data

Cloud resolving LES with 25 Billion grid cells

2002 2009 2015 2016

Views on a system





Mistral Phase 1:

Nodes:

- 1560 x bullx B700 DLC with 2 Intel E5-2680 Haswell@2.5 GHz
- 24 cores/node (36000 total)
- 64/128/256 Gigabyte memory (120 Terabyte total)
- 24 K80 GPUS for remote vis and compute

Performance:

- Peak: 1.4 Petaflop/s
- Application: x 9

Network:

- FatTree with FDR-14 Infiniband (1:2:2 blocking)
- 3 Mellanox SX6536 core 648-port switches

Discs

- 20 Petabyte (usable)

Power consumption

- < 700 kW

Phase 2

Broadwell CPUs

Application x 20

50 Petabyte (usable)

< 1350 kW

Mistral Phase 1:

Nodes:

- 1560 x bullx B700 DLC with 2 Intel E5-2680 Haswell@2.5 GHz
- 24 cores/node (36000 total)
- 64/128/256 Gigabyte memory (120 Terabyte total)
- 24 K80 GPUS for remote vis and compute

Performance:

- Peak: 1.4 Petaflop/s
- Application: x 9

Network:

- FatTree with FDR-14 Infiniband (1:2:2 blocking)
- 3 Mellanox SX6536 core 648-port switches

Discs

- 20 Petabyte (usable)

Power consumption

- < 700 kW

Phase 2

Broadwell CPUs

Application x 20

50 Petabyte (usable)

< 1350 kW

I/O architecture (Phase 1)

Phase 2

50 Petabyte

- Storage capacity: 20 Petabyte
- Lustre 2.5 (+ Seagate patches: some back ports)
- 29 ClusterStor 9000 with 29 Extensions (JBODs)
  - 58 OSS with 116 OST
- ClusterStor 9000 SSUs
  - GridRaid: 41 HDDs, PD-RAID with 8+2(+2 spare blocks)/RAID6, 1 SSD for Log
  - 6 TByte disks
  - SSU: Active/Active failover server pair
  - ClusterStor Manager
  - 1 FDR uplink/server
- Peak performance
  - Infiniband FDR-14: 6.0 GiB/s -> 348 GiB/s
  - CPU/6 GBit SAS: 5.4 GiB/s -> 313 GiB/s
  - 80000 Ops / s
- Multiple metadata servers
  - Root MDS + 4 DNE MDS
  - Active/Active failover (DNEs, Root MDS with Mgmt)

I/O Issues

- No policies (e.g. quota) matching our workflow
  - -> Policy engine: Robin Hood (issues with performance)
- DNE phase 1 does not yet work as expected
  - mv between metadata servers becomes a very slow copy
  - We have to wait for phase 2 (or use backported patches)
- Read cache on the clients does not work as expected
  - due to Lustre consistency data is mostly re-transferred (locking).
  - => Many repeated I/O operations hit disk but that would not be necessary.
- No increase in single stream performance vs previous system (~ 1000 MiB/s)

Tape archive

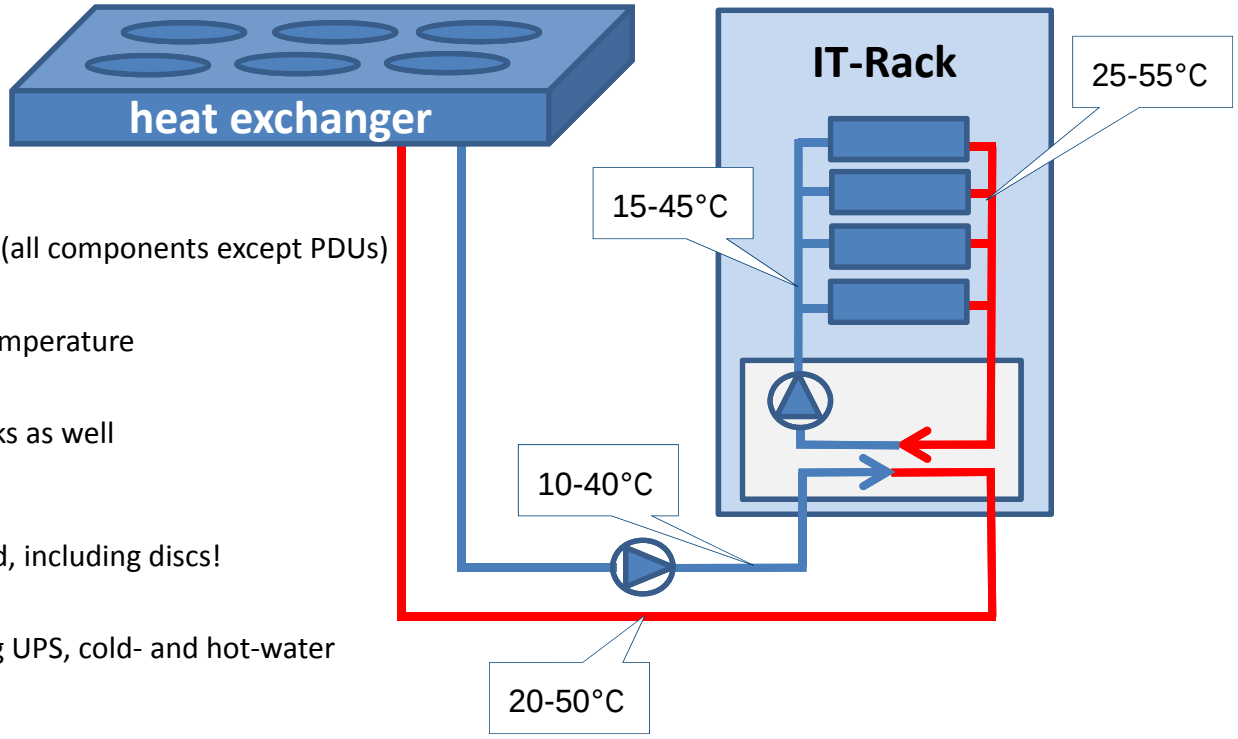
- HPSS has been extended

- Capacity: > 500 PB
- Tentative rate: up to 75 PB/year

- Read/Write peak: 18 GB/s
- Read/Write sustained: 15 GB/s

- Phase 2: oxygene reduction
- Remote double copies of core data (at RZG near Munich)

hot water cooling (DLC)



- Direct liquid cooling of bullx blades (all components except PDUs)
- water-glycol-mix
- Free cooling until ca 40° outside temperature
- Only two cooling circuits
- Water cooled doors for storage-racks as well

Power consumption:

- ~ 500 - 600 kW for typical workload, including discs!
- Phase2: max 1350 kW (contractual)
- PUE: < 1.2 for data centre including UPS, cold- and hot-water cooling (Estimated 1.02 for DLC only)

Benchmarking energy consumption

Mix of the individual benchmarks to simulates mean everyday load on the system

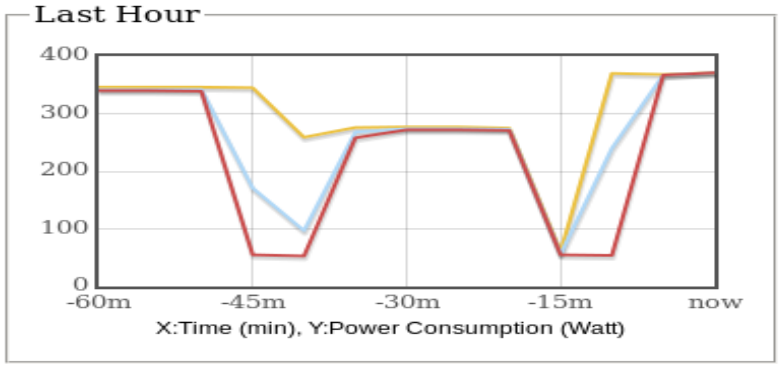
- Fills the whole system. All jobs run concurrently
- Settings and performance (e.g. turbo/non turbo, #cores, SYPD) have to be identical than those used to deliver the performance for individual BMs

This throughput benchmark is used

1. To measure average electrical power
2. To guarantee that no tricks can be played for individual measurements

Measuring energy consumption

- Test: hpcg-2.4 benchmark, 1 node, pure MPI
- Measurements:
  - BMC webinterface:



1.4 GHz

2.5 GHz

2.501GHz (Turbo)

  - sacct report by SLURM
  - srun hdf5 energy-plugin (--acctg-freq=energy=20)

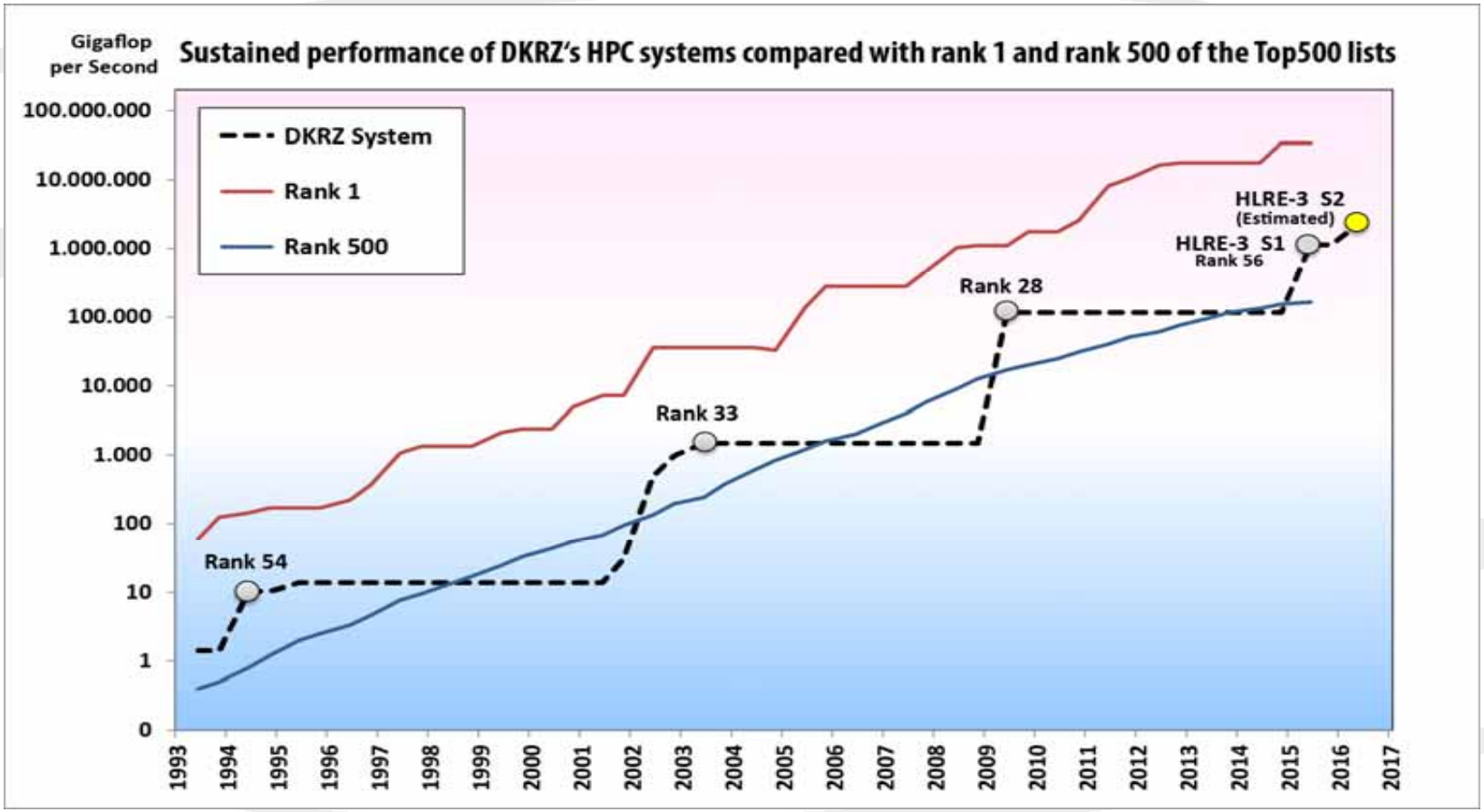
Measuring energy consumption

setup	BMC	sacct	hdf5 plugin
srun 2.5GHz, 24 cores,	346W max	1283 sec elapsed, 434.37kWs Consumed Energy => 338.55W ave	1280 sec elapsed, 58W/347W/ 336W min/max/ave energy
srun 2.501GHz, 24 cores	372W max	1285 sec elapsed, 471.85kWs Consumed Energy => 367.19W ave	1280 sec elapsed, 56W/376W/ 363W min/max/ave energy
srun 1.4GHz, 24 cores	272W ave	1278 sec elapsed, 344.03kWs Consumed Energy => 269.19W ave	1260 sec elapsed, 74W/279W/ 267W min/max/ave energy

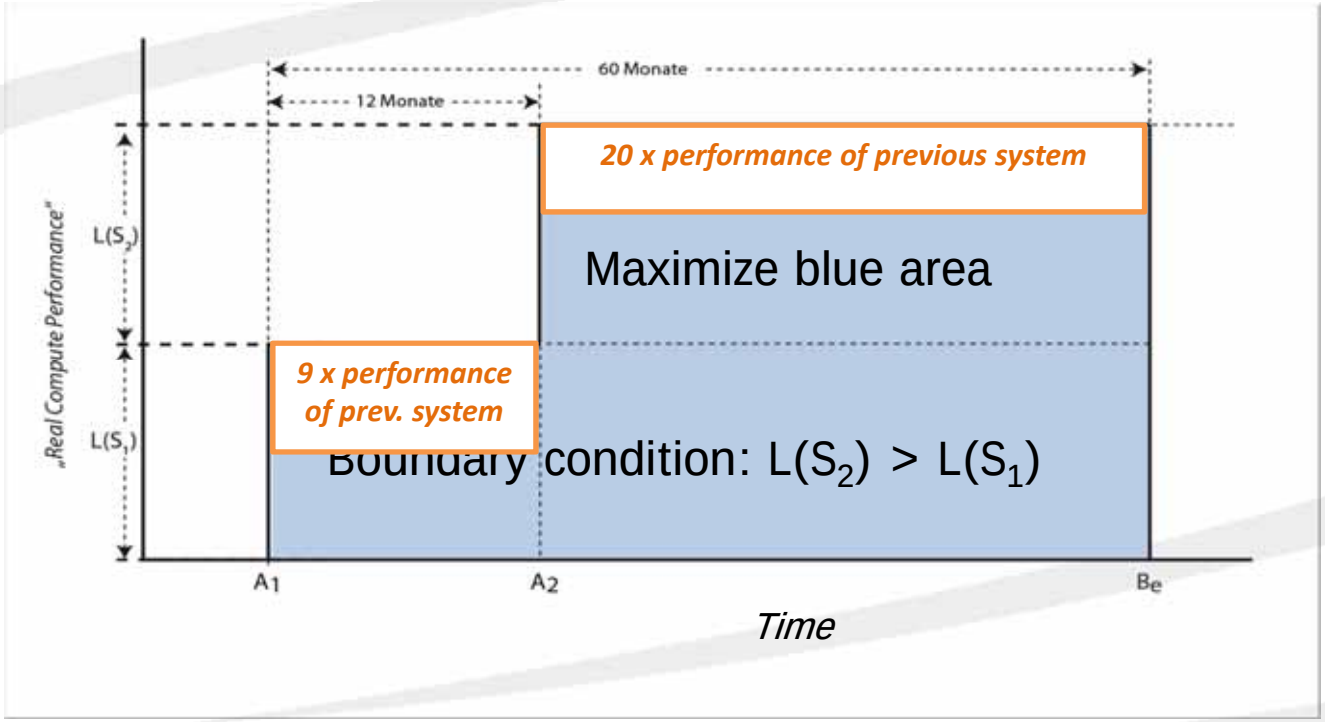
setup	nb of CG solves	ConsumedEnergy	Energy/solve	Equivalent combustion of raw oil
srun 2.5GHz, 24 cores	52	434.37kWs = 434.37e3 J	8.35 kWs	0.0103 kg
srun 2.501GHz, 24 cores	52	471.85kWs = 471.85e3 J	9.07 kWs	0.0112 kg
srun 1.4GHz, 24 cores	50	344.03kWs = 344.03e3 J	6.88 kWs	0.0082 kg

Measuring energy consumption

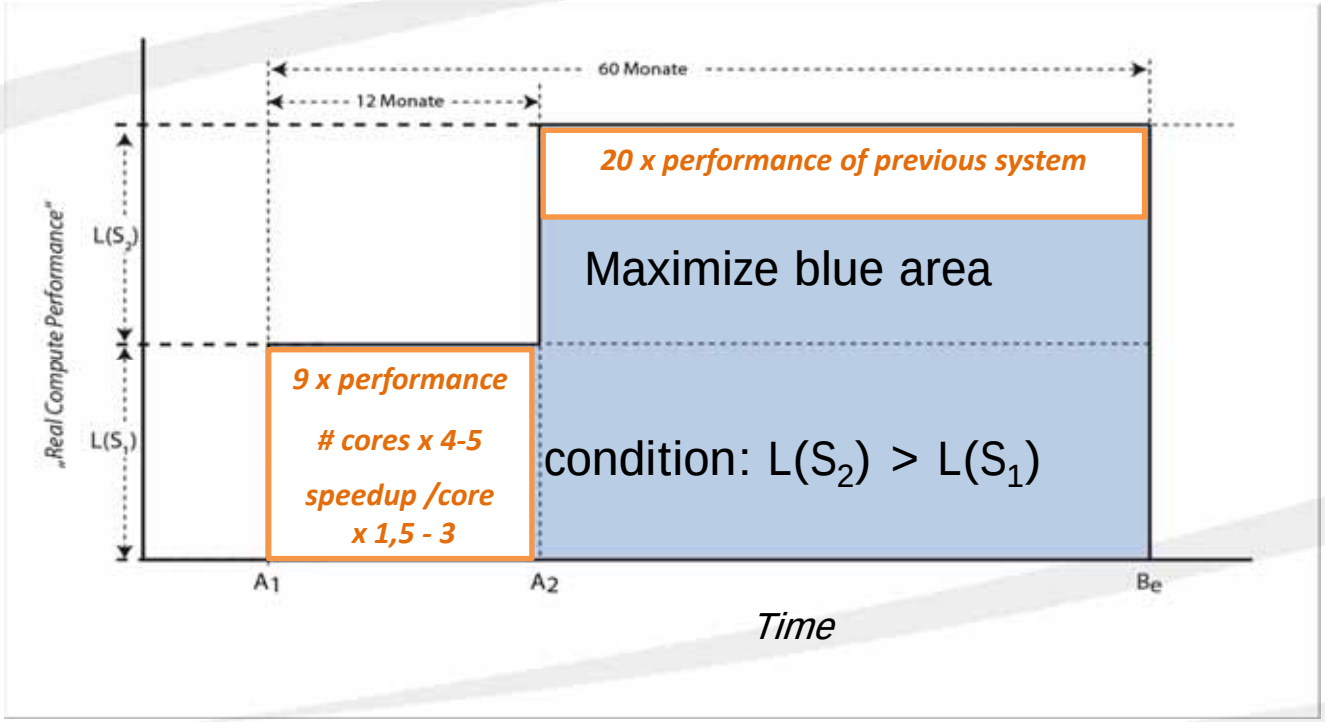
- High Definition Energy Efficiency Monitoring (HDEEM)
  - Cooperation between Bull and Univ. Dresden
  - Motherboards are equipped with FPGAs.
  - Fine grain analysis of data delivered by the Power Management Tools of the bullx Supercomputing Suite.
  - DKRZ has FPGA equipped nodes and can and will do studies using HDEEM


Gigaflop per Second
Sustained performance of DKRZ's HPC systems compared with rank 1 and rank 500 of the Top500 lists
DKRZ System
Rank 1
Rank 500
100.000.000
10.000.000
1.000.000
100.000
10.000
1.000
100
10
1
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Rank 54
Rank 33
Rank 28
HLRE-3 S1
Rank 56
HLRE-3 S2
(Estimated)

How DKRZ defined application performance



How DKRZ defined application performance



How DKRZ defined application performance

A suite of real models selected by user group.

- Configuration (=resolution) as expected to be used in 2015-20
- (but no realistic I/O)

For each: maximal allowed time-to-solution

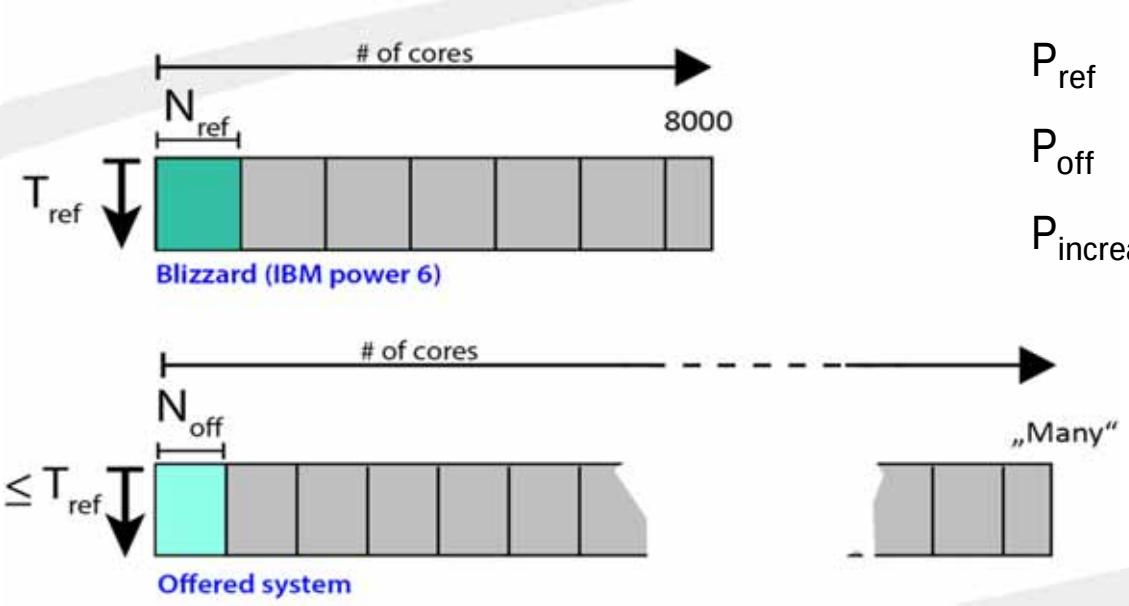
- The number of cores used for to beat this time defined a troughput for this individual BM on the offered system
- A weighted mean of these throughputs is score of the offer

How DKRZ defined application performance

"Real" Application Benchmarks

- ICON global, 20km (N_{ref}: 7872)
- ICON local area 416m (N_{ref}: 4096)
- CCLM (COSMO_RAPS_5.1_CLM) 12 km (N_{ref}: 1024)
- FESOM ocean unstructured grid (N_{ref}: 1024)
- EMAC T42L90, 250 km (N_{ref}: 256)
- MPI-ESM (coupled ESM, T63L95/TP04L40, CMIP5 version) (N_{ref}: 192)
- METRAS (openMP code, meso-scale Atmosphere) (N_{ref}: 32)
- EH6-CDI-PIO (Test for IO server)

How DKRZ defined application performance



$$P_{ref} = 8000 : N_{ref}$$

$$P_{off} = M_{any} : N_{off}$$

$$P_{increase} = P_{off} : N_{ref}$$

Real world issues

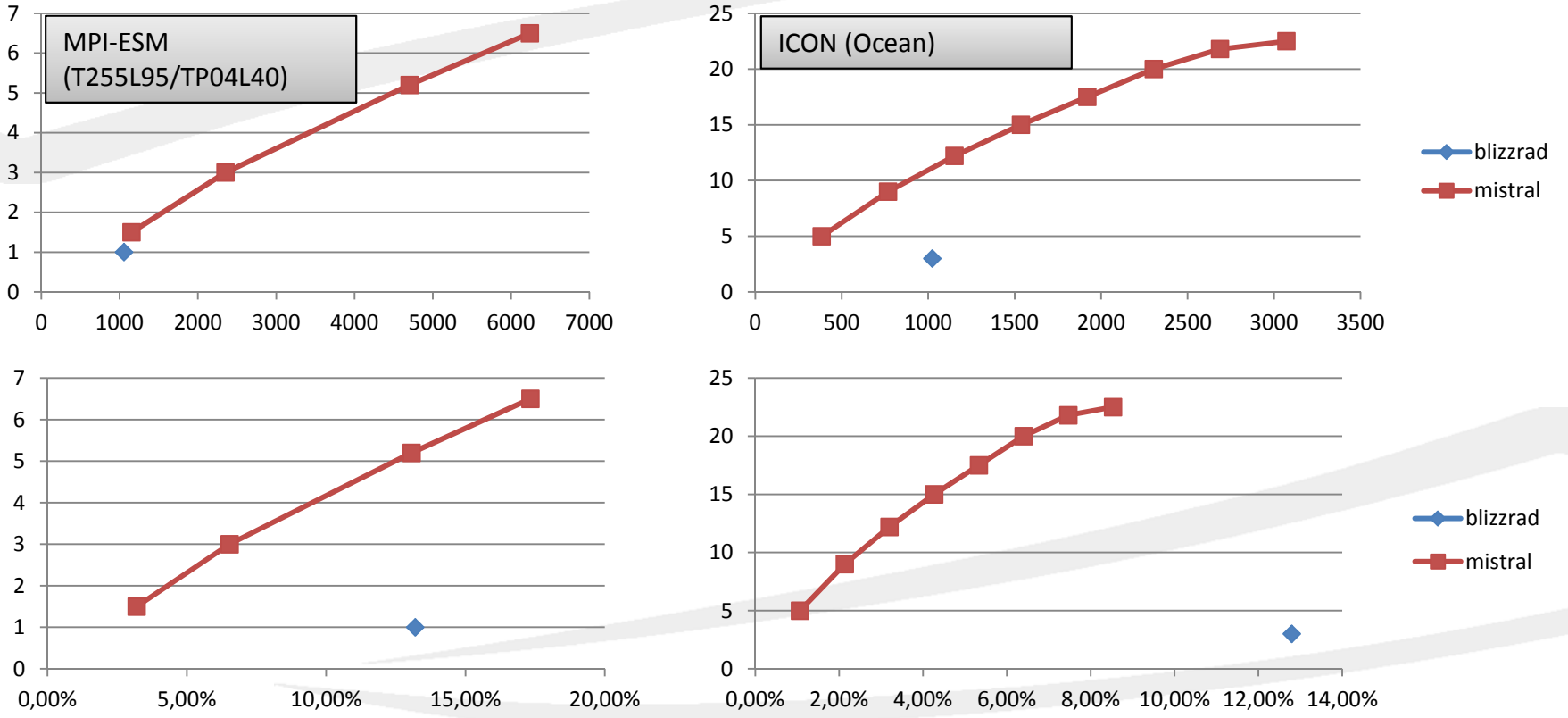
- Which MPI ?
  - bullxMPI, IntelMPI, OpenMPI
  - e.g low node count seems faster with INTEL; high node count faster with bullx
  - „good“ set of flags, env-variables …
  - Initially problems with Intel MPI and > 400 nodes
  - Mellanox libs not supported by Intel MPI
  - compatibiliy: compiler / MPI distribution
- ST vs MT
- Keep the CPU frequency constant (even after reboot)

Real world issues (and support)

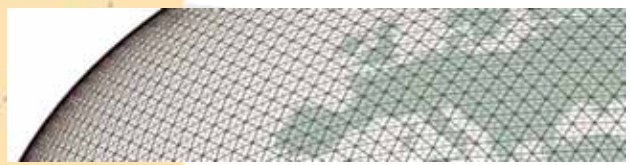
- Which MPI ?
  - bullxMPI, IntelMPI, OpenMPI
  - e.g low node count seems faster with INTEL; high node count faster with bullx
  - „good“ set of flags, env-variables …
  - Initially problems with Intel MPI and > 400 nodes
  - Mellanox libs not supported by Intel MPI
  - compatibiliy: compiler / MPI distribution
- ST vs MT
- Keep the CPU frequency constant (even after reboot)

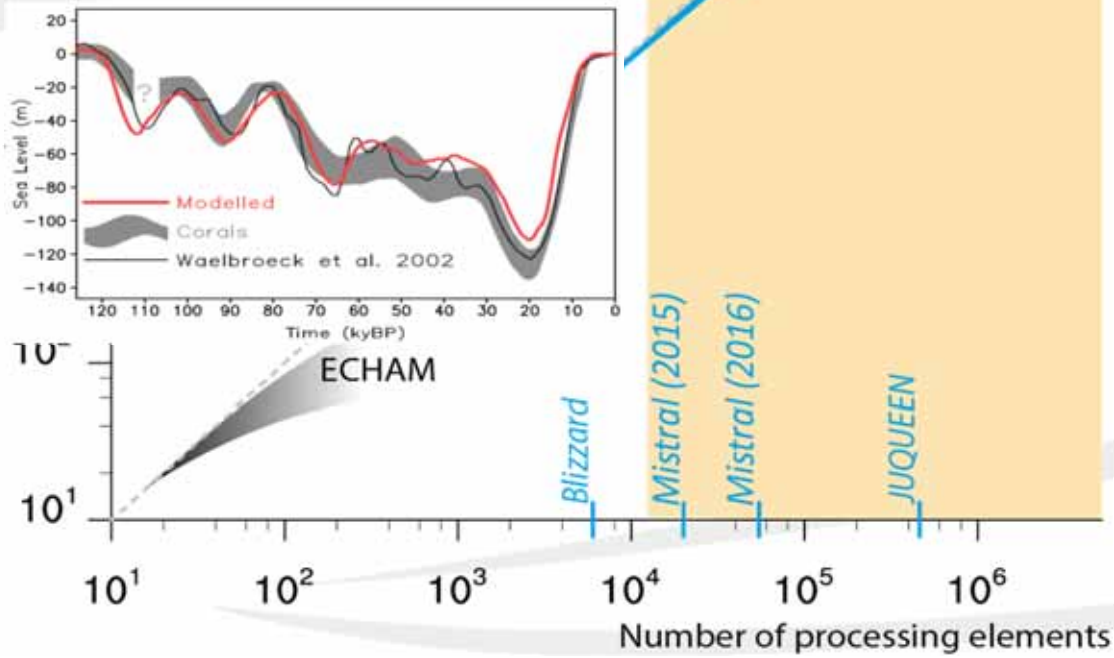
- Bull application support
  - 1 person (in house) to complement DKRZ‘s 2nd and 3rd level support
- Cooperation DKRZ/Bull vs extreme scale computing
  - use case: ICON
  - 2 funded persons (1 nerer to the application, 1 nearer to the system)

Application Performance

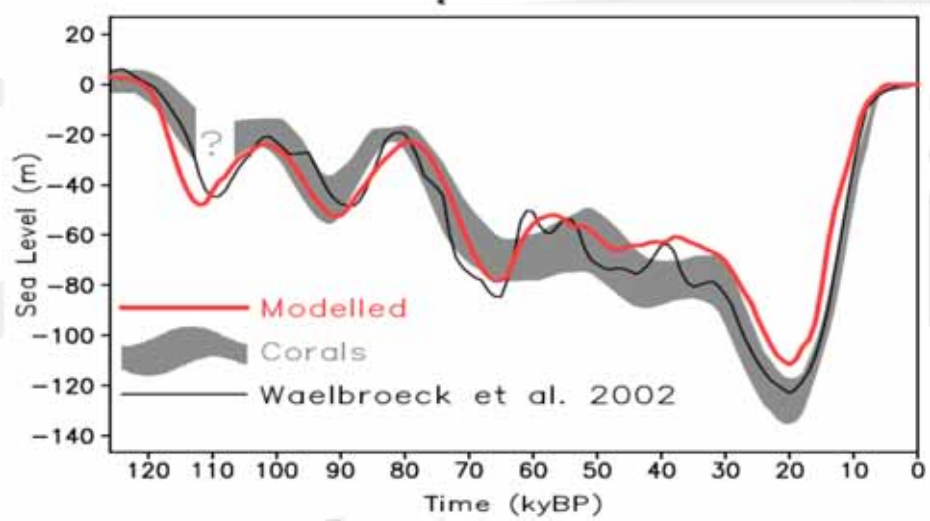


Range of applications: Capacity vs capability





Graph; Bjorn Stevens, internal report HD(CP)2



ECHAM

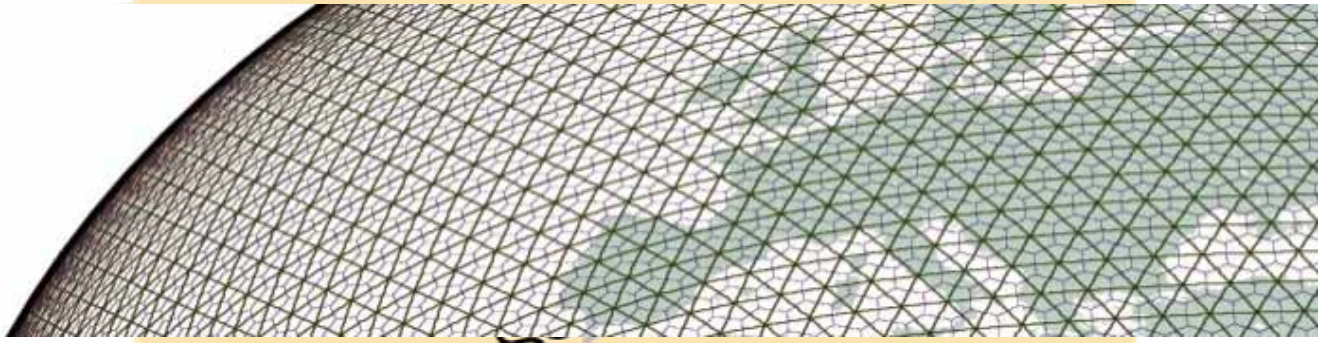
Blizzard

10^1

10^1 10^2 10^3 10

N

PetaFLOPS



Perfect S

ICON

1985: Control Data Cyber-205

- 1 processor
- 0.2 Gigaflops
- 0.03 Gigabyte memory

1988: Cray 2S

- 4 processors
- 2 Gigaflops
- 1 Gigabyte memory

1994: Cray C-916 „Sea"

- 16 processors
- 16 Gigaflops
- 2 Gigabyte memory
- 128 Gigabyte disc space
- 10 Terabyte tape archive

1985 1988 1994 2000

2002: NEC SX-6 „Hurrikan"

- 192 processors
- 1.5 Teraflops
- 1.5 Terabyte memory
- 60 Terabyte disc space
- 3.4 Petabyte tape archive

2009: IBM Power6 „Blizzard"

- 8500 processors
- 158 Teraflops
- 20 Terabyte memory
- 6 Petabyte disc space
- 60 Petabyte tape archive

2015/16: bullx B700 DLC „Mistral"

- 36000 - ca 80000 processors
- 1.5 - 3+ Petaflops
- 120 - 240 Terabyte memory
- 20 - 50 Petabyte disc space
- up to 500 Petabyte tape archive

2002 2009 2015 2016

MISTRAL

ESiWACE

DKRZ

MISTRAL

ESiWACE

EU funded „center of excellence”

DKRZ

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Horizon 2020
European Union funding
for Research & Innovation

The project ESiWACE has received funding (ca 5 Mio €) from the European Union's Horizon 2020 research and innovation programme grant agreement *No 675191*.

Horizon2020 Work Programme 2014-2015, European research infrastructures

Call:	e-Infrastructures
Topic:	EINFRA-5-2015: Centres of excellence for computing applications
Type of action:	Research and Innovation Action

Kick-off will be Dec 1st 2015

esiwace

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

European Commission | Horizon 2020 European Union Funding for Research & Innovation

ESiWACE will

- substantially **improve the efficiency and productivity** of numerical weather and climate simulation on high-performance computing platforms.
- **support the end-to-end workflow** of global Earth system modelling for weather and climate simulation in high performance computing environments.
- foster the **interaction between industry** and the weather and climate community on the exploitation of high-end computing systems, application codes and services.
- **increase competitiveness and growth of the European HPC industry.**

The European **weather and climate science community** will

- drive the governance structure that defines the services to be provided by ESiWACE.

esiwace

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

European Commission | Horizon 2020 European Union Funding for Research & Innovation

ESiWACE will

- substantially **improve the efficiency and productivity** of numerical weather and climate simulation on high-performance computing platforms.
- **support the end-to-end workflow** of global Earth system modelling for weather and climate simulation in high performance computing environments.
- foster the **interaction between industry** and the weather and climate community on the exploitation of high-end computing systems, application codes and services.
- **increase competitiveness and growth of the European HPC industry.**

The European weather and climate science community will

- drive the governance structure that defines the services to be provided by ESiWACE.

esiwace

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Join weather and climate communities

to provide support, training, services

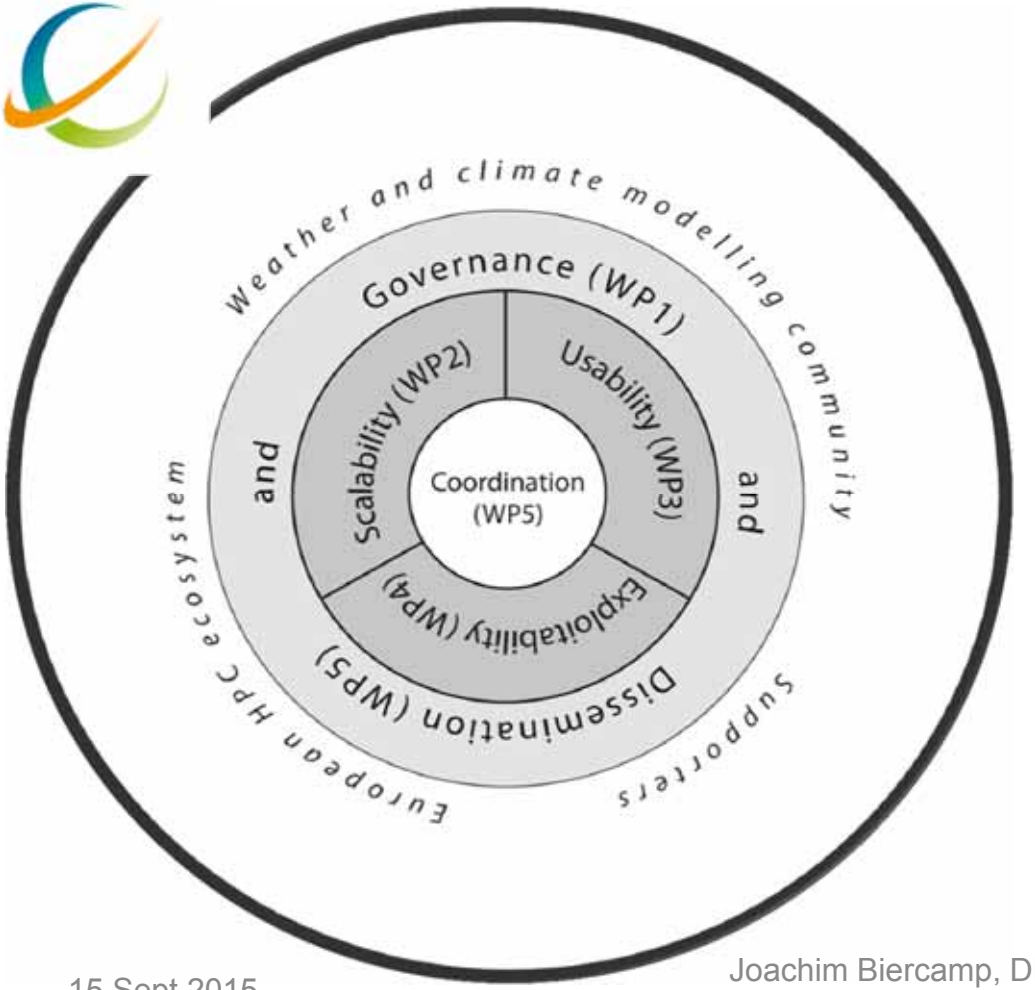
for efficient earth system modelling

using HPC

CO-ORDINATION TEAM

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Nr.	Work Package Title	Lead Institution short name	Co-Lead Institution, short name
WP1	Governance, Engagement & long-term sustainability	CNRS-IPSL, Sylvie Joussaume	DKRZ, Joachim Biercamp
WP2	Scalability	ECMWF, Peter Bauer	CERFACS, Sophie Valcke
WP3	Usability	MPG Reinhard Budich	BSC, Oriol Mula-Valls
WP4	Exploitability	STFC Bryan Lawrence	DKRZ, Thomas Ludwig
WP5	Management & Dissemination	DKRZ, Joachim Biercamp	ECMWF Peter Bauer
Total Amount of Person Months			



WP1 Governance and engagement

- Engagement and governance
- Enhancing community capacity in HPC
- Strategic interaction with HPC ecosystem and HPC industry
- Sustainability and business planning

HPC task force

WP2 Scalability

- Support, training and integration of state of-the-art community models and tools
- Performance analysis and inter-comparisons
- Efficiency enhancement of models and tools
- Preparing for exascale

WP3 Usability

- ESM end-to-end workflows Recommendations
- ESM system software stack recommendations
- ESM scheduling
- Co-Design for Usability

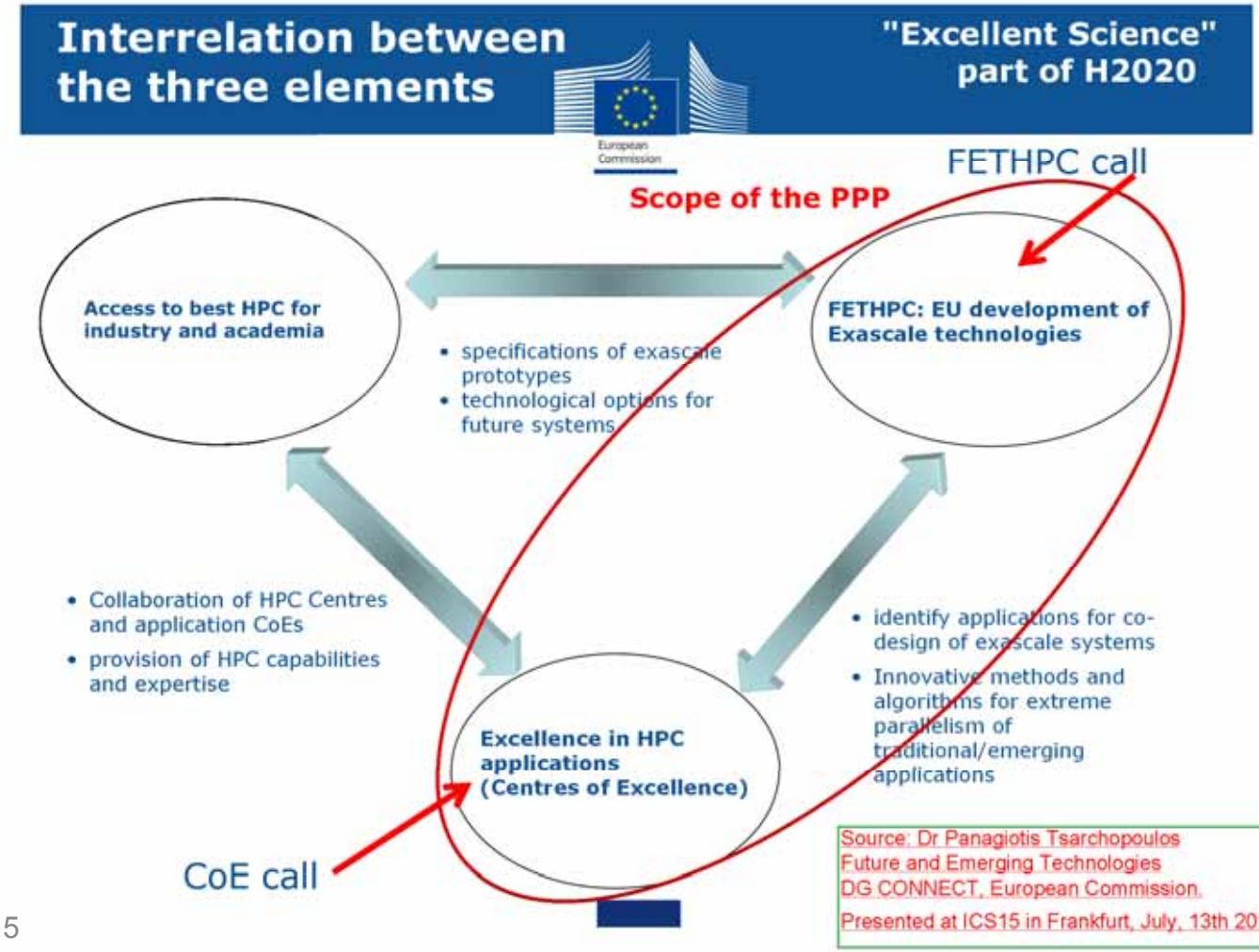
WP4 Exploitability

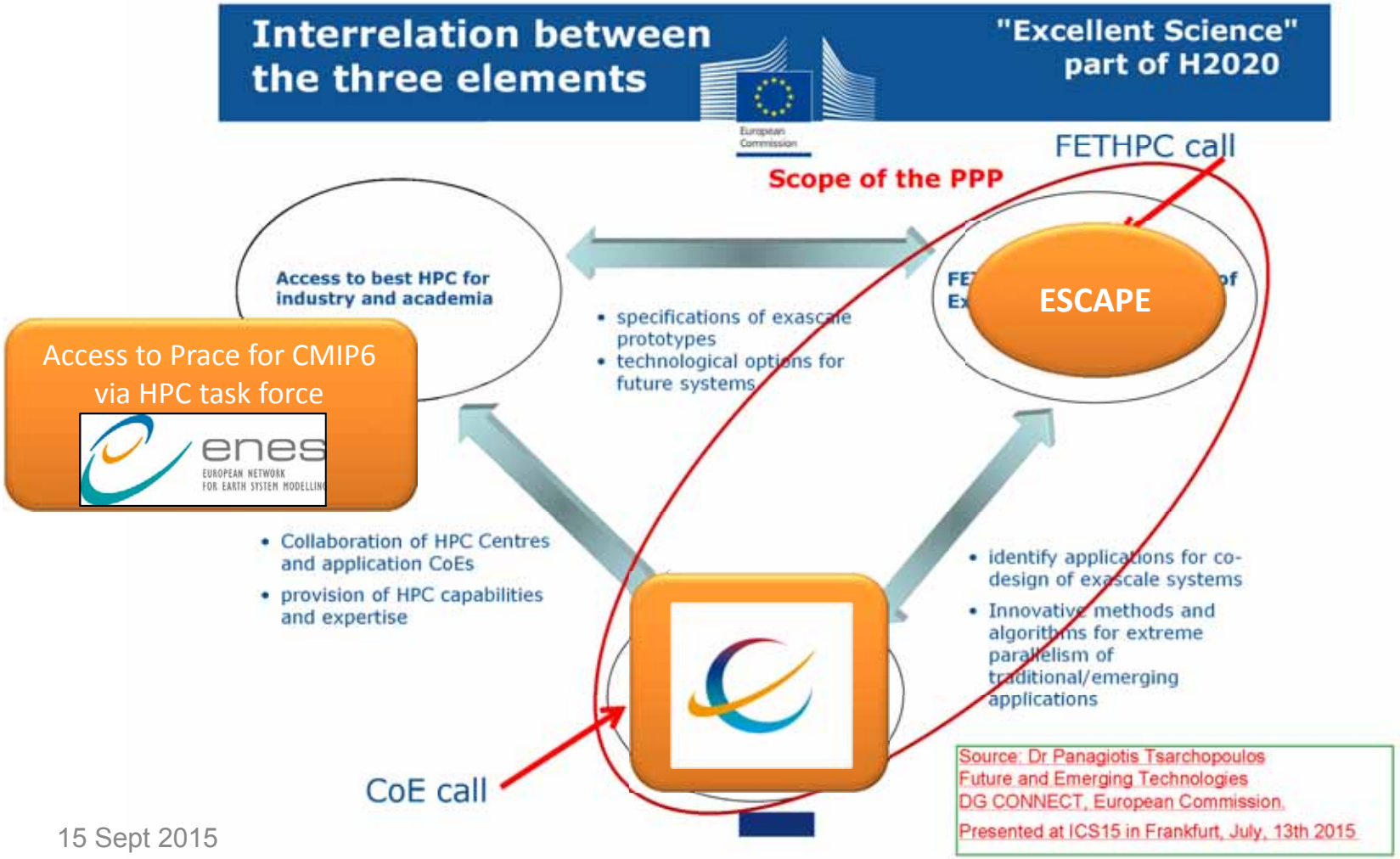
- The business of storing and exploiting high volume climate data
- New storage layout for Earth system data
- New methods of exploiting tape
- Semantic mapping between netCDF and GRIB

WP5 Management and Dissemination



1	Deutsches Klimarechenzentrum GmbH COORDINATOR	DKRZ	Germany
2	European Centre for Medium-Range Weather Forecasts	ECMWF	United Kingdom
3	Centre National de la Recherche Scientifique	CNRS-IPSL	France
4	Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. / Max-Planck-Institut für Meteorologie	MPG	Germany
5	Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique	CERFACS	France
6	Barcelona Supercomputing Center	BSC	Spain
7	Science and Technology Facilities Council	STFC	United Kingdom
8	Met Office	MetO	United Kingdom
9	The University of Reading	UREAD	United Kingdom
10	Sveriges meteorologiska och hydrologiska institut	SMHI	Sweden
11	National University of Ireland Galway (Irish Centre for High End Computing)	ICHEC	Ireland
12	Centro europeo-mediterraneo sui cambiamenti climatici scarl	CMCC	Italy
13	Deutscher Wetterdienst	DWD	Germany
14	Seagate Systems UK Limited	SEAGATE	United Kingdom
15	BULL SAS	BULL	France
16	Allinea Software Limited	ALLINEA	United Kingdom


Interrelation between the three elements
"Excellent Science" part of H2020
European Commission
FETHPC call
Scope of the PPP
Access to best HPC for industry and academia
FETHPC: EU development of Exascale technologies
• specifications of exascale prototypes
• technological options for future systems
• Collaboration of HPC Centres and application CoEs
• provision of HPC capabilities and expertise
• identify applications for co-design of exascale systems
• Innovative methods and algorithms for extreme parallelism of traditional/emerging applications
Excellence in HPC applications (Centres of Excellence)
CoE call
Source: Dr Panagiotis Tsarchopoulos
Future and Emerging Technologies
DG CONNECT, European Commission
Presented at ICS15 in Frankfurt, July, 13th 2015


Interrelation between the three elements
"Excellent Science" part of H2020
European Commission
FETHPC call
Scope of the PPP
Access to best HPC for industry and academia
• specifications of exascale prototypes
• technological options for future systems
ESCAPE
Access to Prace for CMIP6 via HPC task force
enes
EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING
• Collaboration of HPC Centres and application CoEs
• provision of HPC capabilities and expertise
• identify applications for co-design of exascale systems
• Innovative methods and algorithms for extreme parallelism of traditional/emerging applications
CoE call
Source: Dr Panagiotis Tsarchopoulos
Future and Emerging Technologies
DG CONNECT, European Commission
Presented at ICS15 in Frankfurt, July, 13th 2015

SAVE THE DATE

4th ENES HPC Workshop (and ESiWACE GA)

Toulouse, April 6-7 2016

Follow-ons:

climate day at ECMWF workshop, autumn 2017

5th ENES HPC Workshop in Lecce, Italy

Previous workshops:

Lecce 2011, Toulouse 2013, Hamburg 2014