Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

Co-Designing a System for Regional Weather Prediction

Oliver Fuhrer[1]**,** Xavier Lapillonne[1], Guilherme Peretti-Pezzi[2], Carlos Osuna[3], Thomas Schulthess[2,]
(**Philippe Steiner[1]**)

[1]Federal Institute of Meteorology and Climatology MeteoSwiss
[2]Swiss National Supercomputing Centre CSCS, Lugano
[3]Centre for Climate Systems Modeling C2SM, ETH Zurich

.

Current operational system

ECMWF-Model

16 km gridspacing

2 x per day 10 day forecast

COSMO-7

Δx = 6.6 km, Δt = 60 s

393 x 338 x 60 cells

3 x per day 72 h forecast

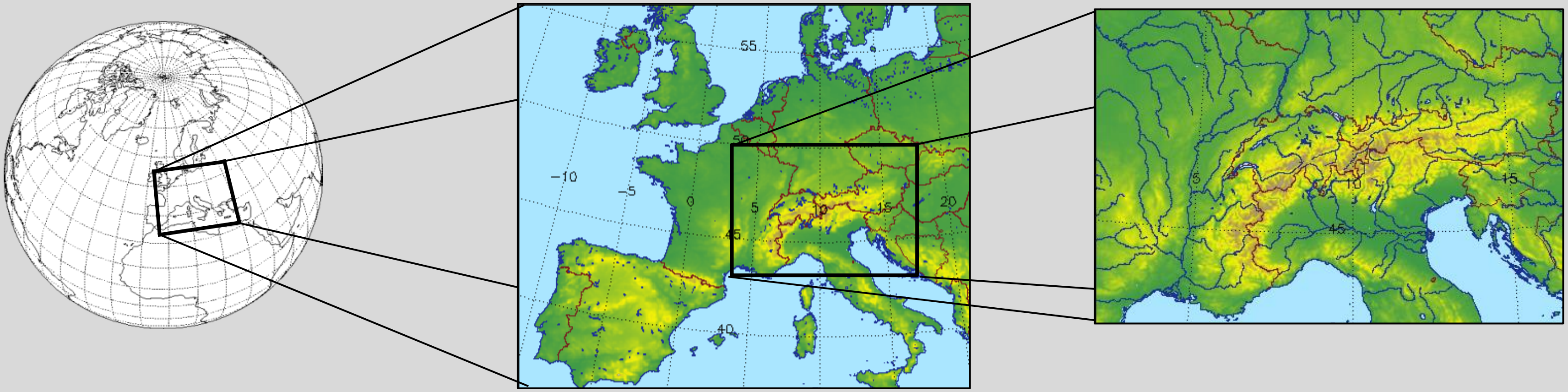
COSMO-2

Δx = 2.2 km, Δt = 20 s

520 x 350 x 60 cells

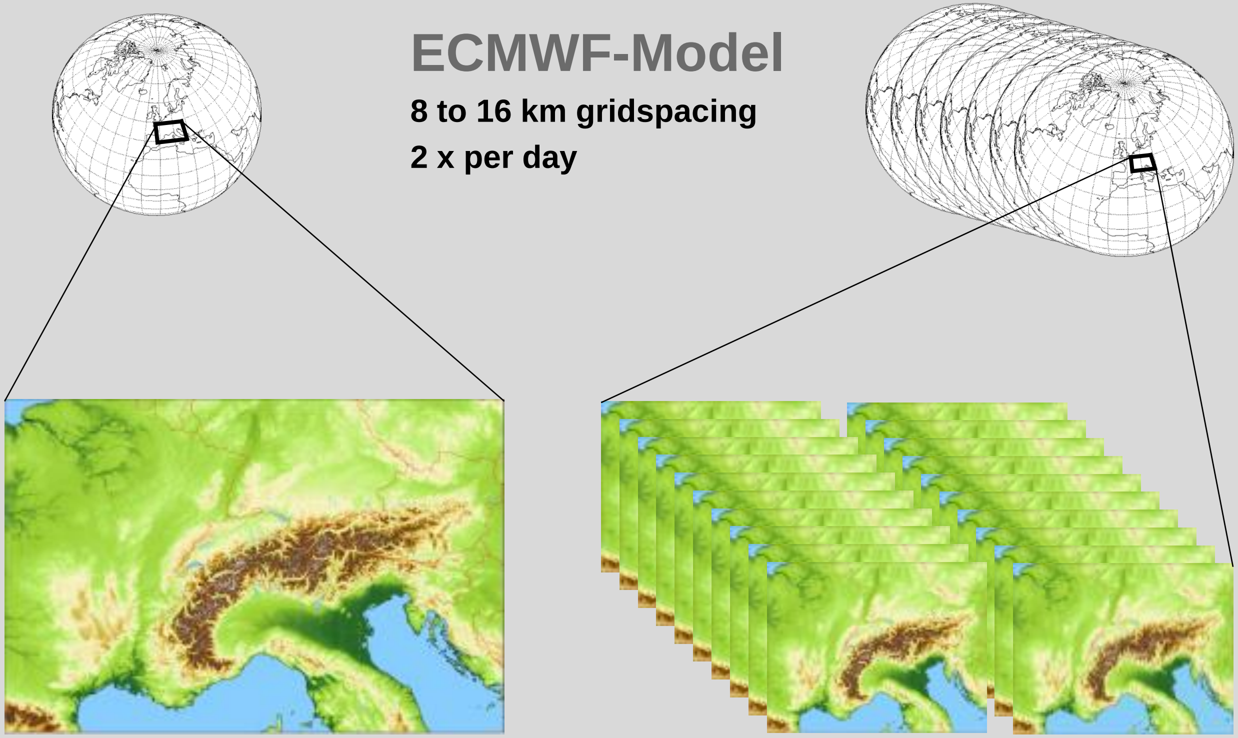
7 x per day 33 h forecast

1 x per day 45 h forecast



Next-generation system

COSMO-1 since 30th September 2015 preoperational



ECMWF-Model

8 to 16 km gridspacing

2 x per day

COSMO-1

Δx = 1.1 km, Δt = 10 s

1158 x 774 x 80 cells

8 x per day:

7 x 33h forecasts

1 x 45h forecast

COSMO-E

Δx = 2.2 km, Δt = 20 s

582 x 390 x 60 cells

2 x per day

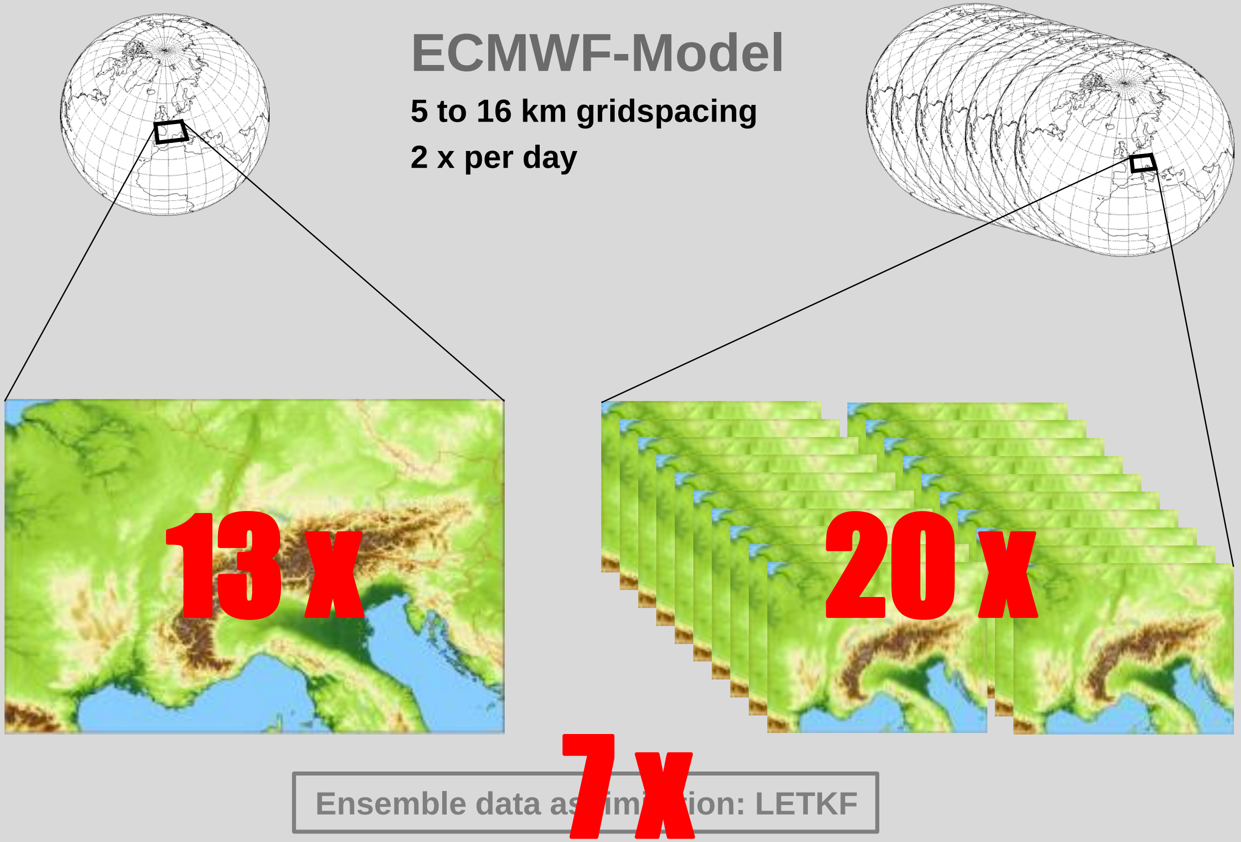
5 d forecast

21 members

Ensemble data assimilation: LETKF

Computational cost = 40 x

(relative to current operational system)



ECMWF-Model

5 to 16 km gridspacing

2 x per day

COSMO-1

1.1 km gridspacing

8 x per day

13 x

20 x

COSMO-E

2.2 km gridspacing

2 x per day

5 d forecast

21 members

7 x

Ensemble data assimilation: LETKF

Production with COSMO @ CSCS

Cray XE6 (Albis/Lema)

MeteoSwiss operational system

Since ~4 years

Next-generation system

Accounting for Moore’s law (factor 4)

Not feasible!

(power, floor space, cost)

Images: CSCS

Co-design: A way out?

Potential

- Time-to-solution driven
- Exclusive usage
- Only one critical application
- Stable configuration (code and system)
- Current code is not optimal
- Novel hardware architectures

Challenges

- Community code
  - Large user base
  - Performance portability
  - Knowhow transfer
- Complex workflow
- High reliability
- Rapidly evolving technology (hardware and software)

Images: CSCS

Co-design: Approach

- Design **software**, **workflow** and **hardware** with the following principles
  - Portability to other users (and hardware)
  - Achieve time-to-solution
  - Optimize energy (and space) requirements
- **Collaborative effort** mainly between
  - MeteoSwiss, C2SM/ETH, CSCS for software since 2010
  - Cray and NVIDIA for new machine since 2013
  - Domain scientists and computer scientists
- Additional funding from the HPCN Strategy (HP2C, PASC)

Images: CSCS

The Swiss Initiative on High-Performance Computing and Networking (HPCN / HP2C)

Passed by Swiss Parliament in 2009

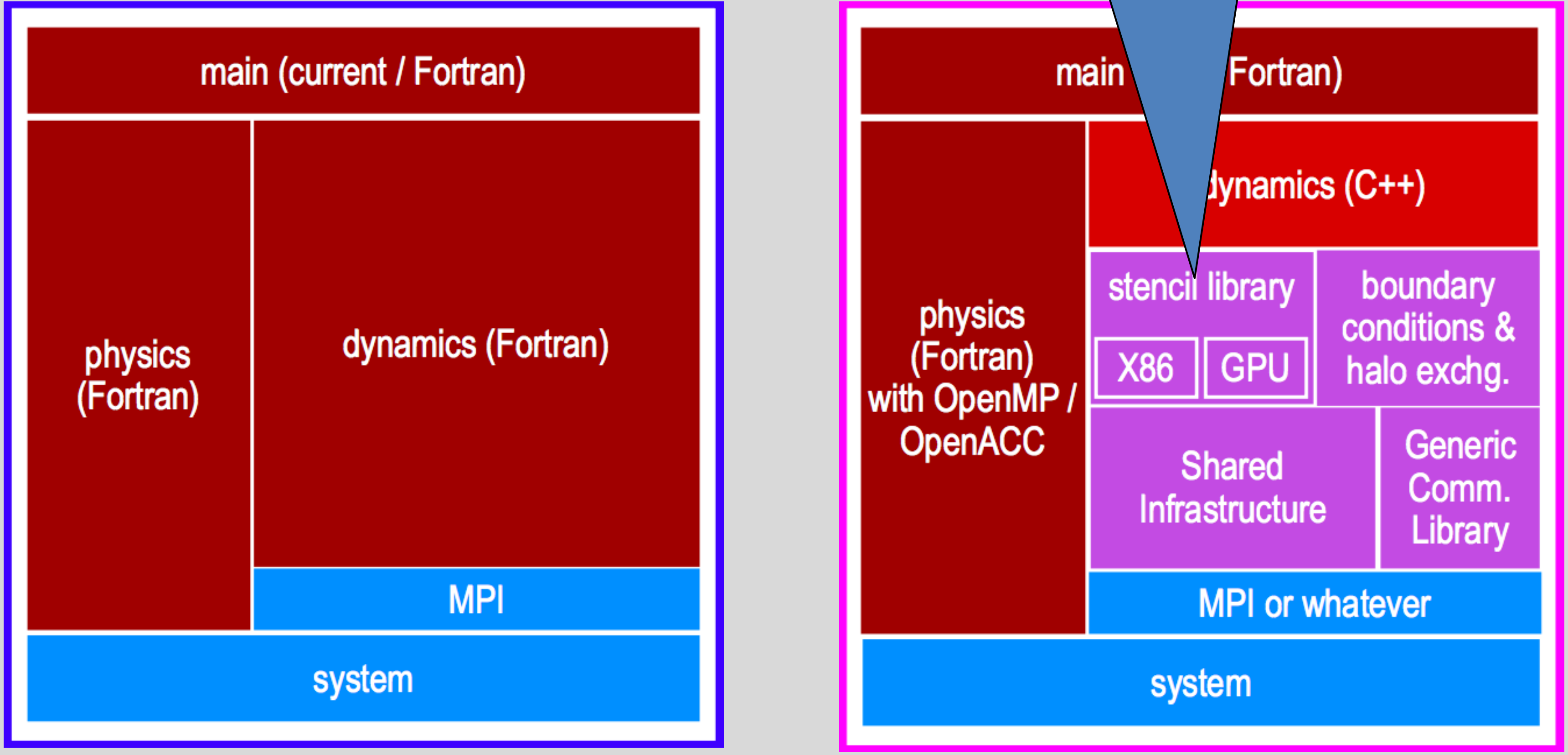
- Investments in
  - new data center in Lugano
  - petascale computing systems
  - application development & know-how (Swiss universities, ETH Zurich/Lausanne)
- Specifically for COSMO
  - support researchers of ETH Zurich
  - software refactoring since fall 2010
  - collaboration MeteoSwiss/C2SM/CSCS
- Development of new MCH system
  - prototype with refactored code since 2013
  - co-designed new machine with NVIDIA & Cray

- New phase **PASC** (Platform for Advanced Scientific Computing) started 2013

Images: CSCS

Current and new code

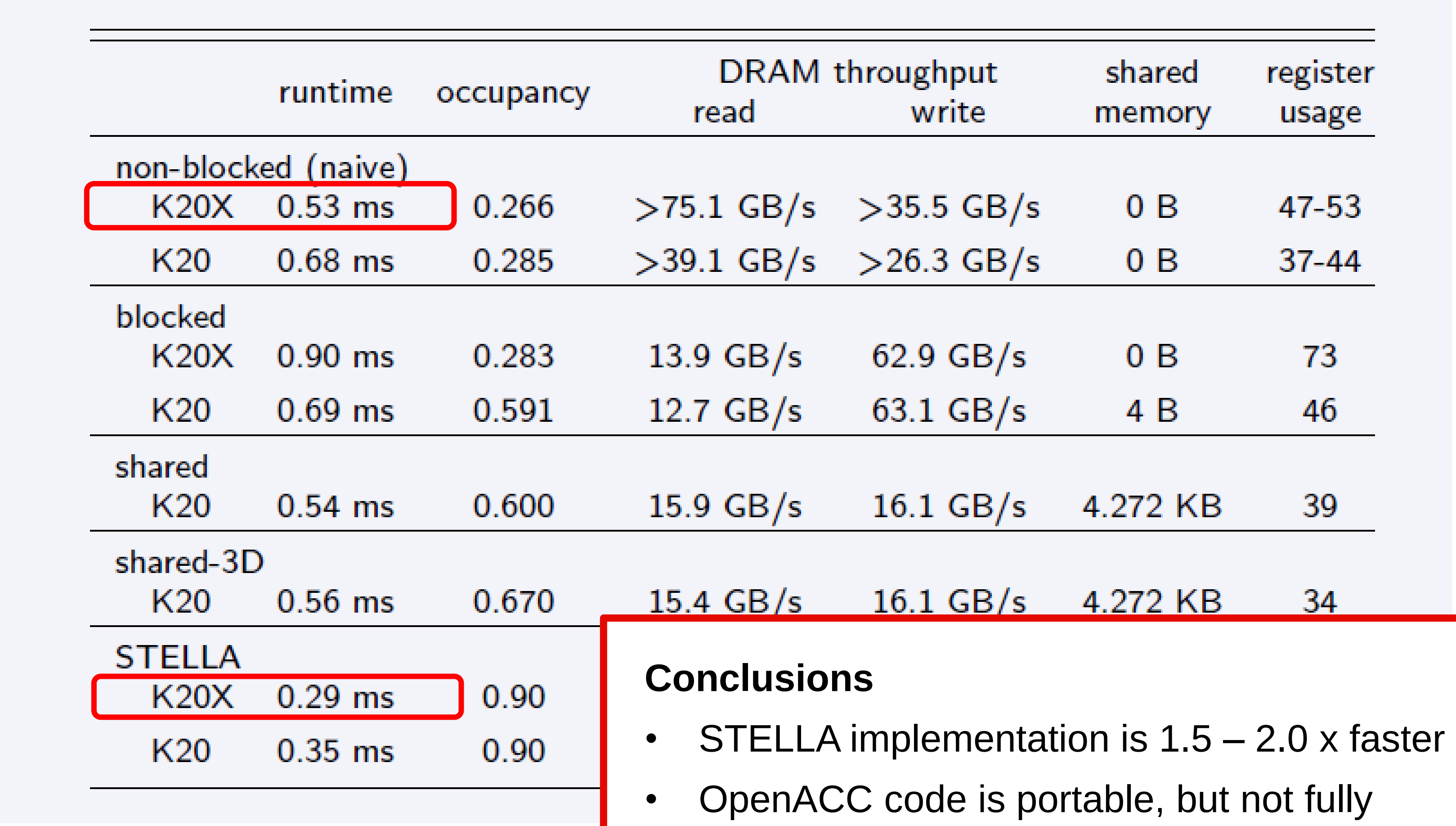
We are currently developing a more general version of STELLA: GridTools (global grids, FEM, …)
>>> see poster



adapted from Fuhrer et al. 2014

OpenACC vs. STELLA

- Comparison using horizontal diffusion
  (also done for vertical advection – not shown)



	runtime	occupancy	DRAM throughput read	DRAM throughput write	shared memory	register usage
non-blocked (naive)						
K20X	0.53 ms	0.266	>75.1 GB/s	>35.5 GB/s	0 B	47-53
K20	0.68 ms	0.285	>39.1 GB/s	>26.3 GB/s	0 B	37-44
blocked						
K20X	0.90 ms	0.283	13.9 GB/s	62.9 GB/s	0 B	73
K20	0.69 ms	0.591	12.7 GB/s	63.1 GB/s	4 B	46
shared						
K20	0.54 ms	0.600	15.9 GB/s	16.1 GB/s	4.272 KB	39
shared-3D						
K20	0.56 ms	0.670	15.4 GB/s	16.1 GB/s	4.272 KB	34
STELLA						
K20X	0.29 ms	0.90				
K20	0.35 ms	0.90				

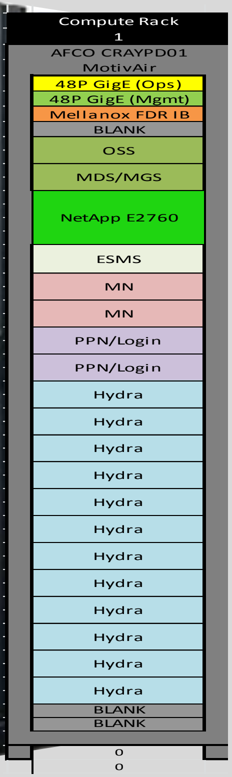
Conclusions

- STELLA implementation is 1.5 – 2.0 x faster
- OpenACC code is portable, but not fully performance portable, many manual optimizations

New MeteoSwiss HPC system

Piz Kesch (Cray CS Storm)

- Installed at CSCS in July 2015
- Public announcement 15th September
- Hybrid system with a mixture of CPUs and GPUs
- “Fat” compute nodes with 2 Intel Xeon E5 2690 (Haswell) and 8 Tesla K80 (each with 2 GK210)
- Only 12 out of 22 possible compute nodes
- Fully redundant (failover for research and development)



Compute Rack 1
AFCO CRAYPD01
MotivAir
48P GigE (Ops)
48P GigE (Mgmt)
Mellanox FDR IB
BLANK
OSS
MDS/MGS
NetApp E2760
ESMS
MN
MN
PPN/Login
PPN/Login
Hydra
Hydra
Hydra
Hydra
Hydra
Hydra
Hydra
Hydra
Hydra
Hydra
Hydra
Hydra
BLANK
BLANK

It is now possible to compare our choice against a more “traditional” choice (e.g. Cray XC40 with Haswell CPUs)

New MeteoSwiss HPC system

Piz Dora (Cray XC40)

- “Traditional” CPU based system
- Compute nodes with 2 Intel Xeon E5-2690 v3 (Haswell)
- Pure compute rack
- Rack has 192 compute nodes
- Very high density (supercomputing line)

Energy Measurement

- We use power clamp for comparison
- Measurements from PMDB and RUR were within 1% of clamp

Piz Dora (Cray XC40)

- **Power clamp**
(external measurement which measures wall consumption including AC/DC conversion, interconnect, but excluding blower)
- 1-2 nodes were down and could not be used (considered in computation)
- **PMDB** (1 Hz, per node)
- **RUR** (total per job)

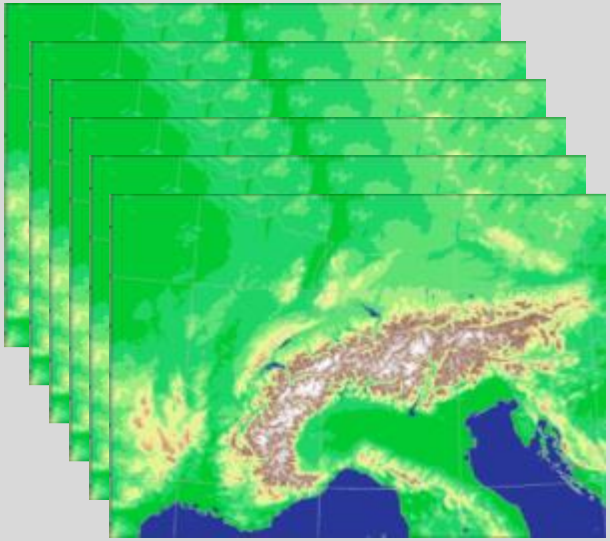
Piz Kesch (Cray CS Storm)

- **Power clamp**
(external measurement which measures wall consumption including AC/DC conversion, interconnect, but excluding blower)
- Other components (mgmt nodes, extra service nodes, drives) powered down

Benchmark

COSMO-E

2.2 km gridspacing
582 x 390 x 60 gridpoints
120 h forecast



Details

- Planned operational setup by MeteoSwiss
- Required time-to-solution = 2h
  (333 ms per timestep)
- Fill a full rack with members
  (keeping sockets per member constant)
- COSMO v5.0
  (with additions for GPU porting and C++ dynamical core)
- Single precision
  (both CPU and GPU not fully optimized)

Results

Note Not sure if this is an apples-to-apples comparison, due to different “character” of systems

	Piz Dora	Piz Kesch	Factor
Sockets @ required time-to-solution for 21 members	~16 CPUs	~7 GPUs	**2.4 x**
Energy per member	6.19 kWh	2.06 kWh	**3.0 x**
Time with 8 sockets per member	13550 s	5980 s	**2.3 x**
Cabinets required to run ensemble at required time-to-solution	0.87	0.39	**2.2 x**

Results Relative to „Old“ Code

(„Old“ = no C++ dycore, double precision)

	Piz Dora (“Old SW”)	Piz Kesch (“New SW”)	Factor
Sockets @ required time-to-solution for 21 members	~26 CPUs	~7 GPUs	**3.7 x**
Energy per member	10.0 kWh	2.06 kWh	**4.8 x**
Time with 8 sockets per member	23075 s	5980 s	**3.8 x**
Cabinets required to run ensemble at required time-to-solution	1.4	0.39	**3.6 x**

„Management summary“

Key ingredients

• Processor performance (Moore’s law)	~2.8 x
• Port to accelerators (GPUs)	~2.3 x
• Code improvement	~1.7 x
• Increase utilization of system	~2.8 x
• Increase in number of sockets	~1.3 x
• Target system architecture to application	

Note
Factor 4x comes from the software refactoring!

Note Solution comes from a combination of investments in hardware, software and workflow

Image: Cray

The Right Performance Metric?

	Piz Dora[1]	Piz Kesch[2]
HPL (TFLOP/s for one full cabinet)	~150	~260
HPCG (TFLOP/s for one full cabinet)	~3.0	~8.1
COSMO (10^9 gridpoint updates per s at scale and time-to-solution)	0.98	2.2

[1] results scaled from benchmark on more cabinets

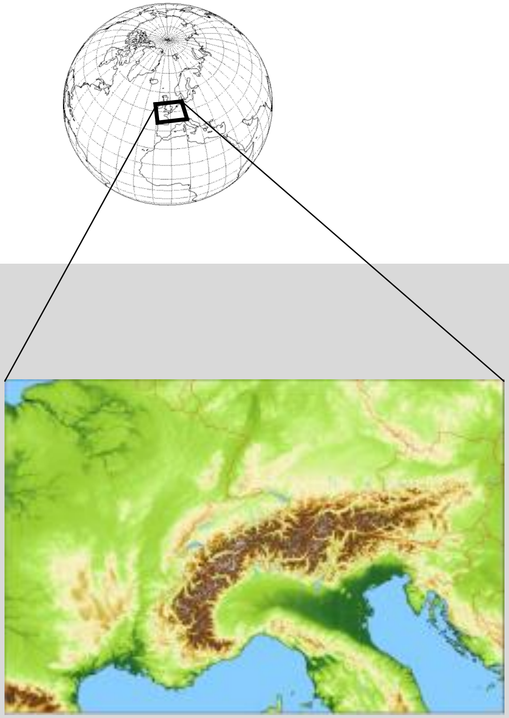
[2] results scaled from 12 to 22 compute nodes per rack

Measurements courtesy of CSCS

Summary

- New forecasting system doubling resolution of deterministic forecast and introducing a convection permitting ensemble
- First element **COSMO-1 preoperational** since 30th September
- Operations of the whole system planed for spring 2016
- **Co-design** (simultaneous code, hardware & workflow re-design) allowed MeteoSwiss to increase operational computational load by 40x within 4–5 years
- New **CS Storm system with fat GPU** nodes since July 2015
- **Energy to solution is a factor 3x smaller** as compared to a “traditional” CPU-based system
- New code to be integrated in **COSMO official version** in 2016

COSMO-1



Thank you for your attention!

Questions?