

Exascale computing: endgame or new beginning for climate modelling

Thomas C. Schulthess



17th Workshop on HPC in Meteorology @ ECMWF, Reading, Wednesday October 26, 2016 T. Schulthess 1

Operational system for NWP at CSCS



Albis & Lema (in production through 3/2016) New system: Kesch & Escha



MeteoSwiss' performance ambitions



We need a 40x improvement between 2012 and 2015 at constant cost



September 15, 2015 **Today's Outlook: GPU-accelerated Weather Forecasting** John Russell

MeteoSwiss New Weather Supercomputer

World's First GPU-Accelerated Weather Forecasting System



2x Racks

48 CPUs

192 Tesla K80 GPUs

> 90% of FLOPS from GPUs Operational in 2016









State of the art implementation of new system for MeteoSwiss

- Albis & Lema: 3 cabinets Cray XE6 installed Q2/2012
- New system needs to be installed Q3/2015
- Assuming 2x improvement in per-socket performance:
 ~20x more X86 sockets would require 30 Cray XC cabinets

New system for Meteo Swiss if we build it like the German Weather Service (DWD) did theirs, or UK Met Office, or ECMWF ... (30 racks XC)

-Current Cray XC30/XC40 platform (space for 40 racks XC)

Thinking inside the box was not a good option!

COSMO-OPCODE: a legacy code migration project





17th Workshop on HPC in Meteorology @ ECMWF, Reading, Wednesday October 26, 2016 T. Schulthess 7

How we solved the problem

Investment in software allowed mathematical improvements and change in architecture



There is no silver bullet!





Why commodity processors?

\$500,000,000

\$2,000,000,000

\$13,000

Source: Andy Keane @ ISC'10



25 Years CSCS, Lugano, Wednesday October 19, 2016 T. Schulthess 9





The end of Denard Scaling



p substrate, doping $\alpha^* N_A$

Source: Ronald Luijten, IBM-ZRL

Robert H. Dennard (1974)



CONSEQUENCE: Higher density: $\sim \alpha^2$ Higher speed: $\sim \alpha$ Power/ckt: $\sim 1/\alpha^2$ Power density: \sim constant





Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

Who consumes how much energy on a 28nm processor

•64 bit floating point unit: 20 pJ
•256-bit access 8kB SRAM: 50 pJ
•256-bit bus across die: 1,000 pJ
•Read/write to DRAM: 16,000 pJ

By a wide margin, most energy is spend in moving data on the die and to memory

Developing algorithms that maximise data locality should be THE TOP PRIORITY

Source: Bill Dally, 2011

17th Workshop on HPC in Meteorology @ ECMWF, Reading, Wednesday October 26, 2016 T. Schulthess 13

Moore's Law 2008-2020 Semiconductor Device Scaling Factors

Technology	45000	00					
(High Volume)	(2008)	32nm (2010)	22nm (2012)	16nm	11nm	8nm	5nm
Transistor density	175	4 75	(2012)	(2014)	(2016)	(2018)	(2020)
Frequencyscaling	1.70	1.75	1.75	1.75	1.75	1.75	1.75
Voltore (LLL)	15%	10%	8%	5%	406	20/	-
voltage (Vdd) scaling	-10%	-7.5%	50/	0.50	470	370	2%
Dimension & Capacitance	0.75		-370	-2.5%	-1.5%	-1%	-0.5%
SD Leakage cooling to t	0./5	0.75	0.75	0.75	0.75	0.75	0.75
ob Leakage scaling/micron	1X Optimistic to 1,43X Pessimistic						

Sources: International Technology Roadmap for Semiconductors and Intel

Moore's Law Takes Miracles ... But It Isn't The Miracle That Will Carry The Day

Source: Rajeeb Hazra's (HPC@Intel) talk at SOS14, March 2010

"Piz Daint"

Cray XC30 with 5272 hybrid, GPU accelerated compute nodes

Compute node:

∠2.5 GHz (~0.38 ns)

~0.73 GHz (~1.4ns)

- > Host: Intel Xeon E5 2670 (SandyBridge 8c) 4
- > Accelerator: One NVIDIA K20X GPU (GK110)

Architectural diversity is here to stay, because it is a consequence of the dusk of CMOS scaling (Moore's Law)

What are the implications?

Complexity in software is one, but we don't understand all implications

Physics of the computer matters more than ever

$$\begin{cases} \frac{\partial u}{\partial t} = \left[-\left\{ \frac{1}{a \cos \varphi} \frac{\partial E_h}{\partial \lambda} - vV_a \right\} - \left[\frac{\dot{c}}{\partial U} \right] \left[-\frac{1}{\rho a \cos \varphi} \left(\frac{\partial t'}{\partial \lambda} - \frac{1}{\sqrt{7}} \frac{\partial p_0}{\partial \lambda} \frac{\partial t'}{\partial \zeta} \right) + M_u \right] \\ \frac{\partial v}{\partial t} = \left[-\left\{ \frac{1}{a} \frac{\partial E_h}{\partial \varphi} + uV_a \right\} \right] - \left[\frac{\dot{c}}{\partial U} \right] \left[-\frac{1}{\rho a} \left(\frac{\partial p'}{\partial \varphi} - \frac{1}{\sqrt{7}} \frac{\partial p_0}{\partial \varphi} \frac{\partial t'}{\partial \zeta} \right) + M_u \right] \\ \frac{\partial w}{\partial t} = \left[-\left\{ \frac{1}{a \cos \varphi} \left(u \frac{\partial w}{\partial \lambda} + v \cos \varphi \frac{\partial w}{\partial \varphi} \right) \right\} - \left[\frac{\dot{c}}{\partial U} \right] \left[-\frac{1}{\sqrt{7}} \frac{\partial p_0}{\partial \varphi} \frac{\partial t'}{\partial \zeta} \right] + M_u + g \frac{\rho_0}{\rho} \left\{ \frac{(T - T_0)}{T} - \frac{T_0 p'}{T p_0} + \left(\frac{R_v}{R_d} - 1 \right) q^v - q^l - q^l \right\} \right] \\ pressure \frac{\partial p'}{\partial t} = \left[-\left\{ \frac{1}{a \cos \varphi} \left(u \frac{\partial p'}{\partial \lambda} + v \cos \varphi \frac{\partial p'}{\partial \varphi} \right) \right\} - \left[\frac{\dot{c}}{\partial \zeta} \right] \left[\frac{d p'}{\partial \zeta} + g \rho_0 w - \frac{c_{pd}}{c_{vd}} p D \right] \right] \\ temperature \frac{\partial T}{\partial t} = \left[-\left\{ \frac{1}{a \cos \varphi} \left(u \frac{\partial T}{\partial \lambda} + v \cos \varphi \frac{\partial q^v}{\partial \varphi} \right) \right\} - \left[\frac{\dot{c}}{\partial \zeta} \right] \left[\frac{1}{\rho c_{vd}} p D \right] + Q_T \right] \\ water \frac{\partial q^v}{\partial t} = \left[-\left\{ \frac{1}{a \cos \varphi} \left(u \frac{\partial q^{l'J}}{\partial \lambda} + v \cos \varphi \frac{\partial q^v}{\partial \varphi} \right) \right\} - \left[\frac{\dot{c}}{\partial \zeta} \right] \left[\frac{g}{\sqrt{7}} \frac{\rho}{\rho} \frac{\partial P_{l,f}}{\partial \zeta} + S^{l,f} + M_{q^l,f} \right] \\ turbulence \frac{\partial e_t}{\partial t} = \left[-\left\{ \frac{1}{a \cos \varphi} \left(u \frac{\partial q^{l'J}}{\partial \lambda} + v \cos \varphi \frac{\partial q^l}{\partial \varphi} \right) \right\} - \left[\frac{\dot{c}}{\partial \zeta} \right] + K_m^v \frac{g \rho_0}{\sqrt{7}} \left\{ \left(\frac{\partial u}{\partial \zeta} \right)^2 + \left(\frac{\partial v}{\partial \zeta} \right)^2 \right\} + \frac{g}{\rho \theta_v} F^{\theta_v} - \frac{\sqrt{2}e_t^{3/2}}{\alpha_{Ml}^2} + M_{e_t} \right] \\ \frac{Timestep}{turbulence} \frac{\partial e_t}{\partial t} = \left[\frac{1}{\left\{ a \cos \varphi \left(u \frac{\partial e_t}{\partial \lambda} + v \cos \varphi \frac{\partial e_t}{\partial \varphi} \right\} \right\} - \left[\frac{\dot{c}}{\partial \zeta} \right] + K_m^v \frac{g \rho_0}{\sqrt{7}} \left\{ \left(\frac{\partial u}{\partial \zeta} \right)^2 + \left(\frac{\partial v}{\partial \zeta} \right)^2 \right\} + \frac{g}{\rho \theta_v} F^{\theta_v} - \frac{\sqrt{2}e_t^{3/2}}{\alpha_{Ml}^2} + M_{e_t} \right] \\ \frac{d v}{d t} = \left[\frac{1}{\left\{ a \cos \varphi \left(u \frac{\partial e_t}{\partial \lambda} + v \cos \varphi \frac{\partial e_t}{\partial \varphi} \right\} \right\} - \left[\frac{\partial e_t}{\partial \zeta} \right] + K_m^v \frac{g \rho_0}{\sqrt{7}} \left\{ \left(\frac{\partial u}{\partial \zeta} \right\} + \left(\frac{\partial v}{\partial \zeta} \right)^2 \right\} + \frac{g}{\rho \theta_v} F^{\theta_v} - \frac{\sqrt{2}e_t^{3/2}}{\alpha_{Ml}^2} + M_{e_t} \right] \\ \frac{d v}{\partial t} = \left[\frac{\partial u}{\partial t} + v \cos \varphi \frac{\partial v}{\partial \varphi} \right] - \left[\frac{\partial u}{\partial \zeta} \right] + \left[\frac{\partial$$

Stencil example: Laplace operator in 2D

Two main components of an operator on a structured grid
1. Loop-logic defines stencil application domain and order
2. Stencil defines the operator to be applied

do k = kstart, kend
do j = jstart, jend
do i = istart, iend
lap(i, j, k) = -4.0 * data(i, j, k) + &
data(i+1, j, , k) + data(i-1, j , k) + &
data(i , j+1, k) + data(i , j-1, k)
end do
end do
end do


```
enum { data, lap };
template<typename TEnv>
struct Laplace
{
  STENCIL STAGE(Tenv)
  STAGE PARAMETER(FullDomain, data)
  STAGE PARAMETER(FullDomain, lap)
  static void Do()
  {
    lap::Center() =
      -4.0 * data::Center() +
      data::At(iplus1) +
      data::At(iminus1) +
      data::At(jplus1) +
      data::At(jminus1);
  }
};
```

```
IJKRealField lapfield, datafield;
Stencil stencil;
StencilCompiler::Build(
pack parameters(
    Param<lap, cInOut>(lapfield),
    Param<data, cIn>(datafield)
),
  concatenate sweeps(
    define sweep<KLoopFullDomain>(
      define_stages(
        StencilStage<Laplace, IJRangeComplete>()
);
stencil.Apply();
```


Stencil

```
enum { data, lap };
template<typename TEnv>
struct Laplace
{
  STENCIL STAGE(Tenv)
  STAGE PARAMETER(FullDomain, data)
  STAGE PARAMETER(FullDomain, lap)
  static void Do()
  {
    lap::Center() =
      -4.0 * data::Center() +
      data::At(iplus1) +
      data::At(iminus1) +
      data::At(jplus1) +
      data::At(jminus1);
 }
};
```

Loop logic

```
IJKRealField lapfield, datafield;
Stencil stencil;
```

```
StencilCompiler::Build(
```

```
pack_parameters(
```

),

```
Param<lap, cInOut>(lapfield),
```

```
Param<data, cIn>(datafield)
```

```
concatenate_sweeps(
    define_sweep<KLoopFullDomain>(
        define_stages(
            StencilStage<Laplace, IJRangeComplete>()
        )
      )
    )
    )
stencil.Apply();
```


Architecture dependent backend

- The same user-level code can be compiled with different, architecture dependent backends
- multi-core CPU (x86) SIMD
 - kij-storage
 - ij-blocking
 - Coarse: OpenMP theads
 - Fine: vectorisation by compiler

• GPU (Tesla) – SIMT

- ijk-storage
- Coarse: CUDA thread blocks
- Fine: CUDA threads
- software managed caching

References and Collaborators

- Peter Messmer and his team at the NVIDIA co-design lab at ETH Zurich
- Teams at CSCS and Meteo Suisse, group of Christoph Schaer @ ETH Zurich
- O. Fuhrer, C. Osuna, X. Lapillonne, T. Gysi, B. Cumming, M. Bianco, A. Arteaga, T. C. Schulthess, "Towards a performance portable, architecture agnostic implementation strategy for weather and climate models", Supercomputing Frontiers and Innovations, vol. 1, no. 1 (2014), see <u>superfri.org</u>
- G. Fourestey, B. Cumming, L. Gilly, and T. C. Schulthess, "First experience with validating and using the Cray power management database tool", Proceedings of the Cray Users Group 2014 (CUG14) (see <u>arxiv.org</u> for reprint)
- B. Cumming, G. Fourestey, T. Gysi, O. Fuhrer, M. Fatica, and T. C. Schulthess, "Application centric energy-efficiency study of distributed multi-core and hybrid CPU-GPU systems", Proceedings of the International Conference on High-Performance Computing, Networking, Storage and Analysis, SC'14, New York, NY, USA (2014). ACM
- T. Gysi, C. Osuna, O. Fuhrer, M. Bianco and T. C. Schulthess, "STELLA: A domain-specific tool for structure grid methods in weather and climate models", to be published in Proceedings of the International Conference on High-Performance Computing, Networking, Storage and Analysis, SC'15, New York, NY, USA (2015). ACM

COSMO: old and new (refactored) code

GTC Europe, Amsterdam, Thursday September 29, 2016 T. Schulthess 30

ational Supercomputing Centre

The good news

C++ standard is evolving quickly and implementations follow!

C++ 11, 14, HPX-3, ... 17, 20, ...

Who will pay for the implementation of Fortran, OpenACC, OpenMP, ...?

NVIDA DGX-1 WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER 170TF | "250 servers in-a-box" | nvidia.com/dgx1

\$129,000 for 8 GPUs, or \$16k a piece

\$500,000,000 \$2,000,000,000 \$13,000

Source: Andy Keane @ ISC'10

Lugano

Switzerland

Platform for Advanced Scientific Computing Conference

26-28 June 2017

Call for Papers

The Platform for Advanced Scientific Computing (PASC) invites submissions for the PASC17 Conference, co-sponsored by the Association for Computing Machinery (ACM) and SIGHPC, which will be held at the Palazzo dei Congressi in Lugano, Switzerland, from June 26 to 28, 2017.

Luqano

Switzerland

Platform for Advanced Scientific Computing Conference

26-28 June 2017

PASC17 is an interdisciplinary event in high performance computing that brings together domain science, applied mathematics and computer science - where computer science is focused on enabling the realization of scientific computation.

We are soliciting high-quality contributions of original research relating to high performance computing in eight domain-specific tracks:

Areas of interest include (but are not limited to):

- The use of advanced computing systems for large-scale scientific applications

- Implementation strategies for science applications in energy-efficient computing architectures
- Domain-specific, languages, libraries or frameworks
- The integration of large-scale experimental and observational scientific data and high-IN THREE - DIMENISIONAL CARTERIAN performance data analytics and computing
- Best practices for sustainable software development and scientific application development

Committee Chairs

Jack Wells (Oak Ridge National Laboratory, USA) Torsten Hoefler (ETH Zurich, Switzerland)

Submission Guidelines

initialize sun and s We invite papers of 5-10 pages in length, which will be reviewed double blind. Full submissions guidelines can be found at www.pasc17.org. while of het assigned do

- Submissions close: 12 December 2016
- First review notification: 31 January 2017
- Revised submissions close: 1 March 2017
- Final acceptance notification: 11 April 2017

Conference Participation and Proceedings

Accepted manuscripts will be **published in the ACM Digital Library** on the first day of the conference. Authors will be given 30-minute presentation slots at the conference, grouped in topically focused, parallel sessions.

Post-Conference Journal Submission

Following the conference, authors will have the opportunity to develop their papers for publication in a relevant, computationally focused, domain-specific journal.

Authors thus stand to benefit from the rapid and broad dissemination of results afforded by the conference venue and associated proceedings, and, from the impact associated with publication in a high-quality scientific journal. (X1, X2, X3)= (X, 4, 7) and

pasc17.pasc-conference.org

How = Ki + Mis end if f (Unow)/f(xi) > = then end for

METROPOLIS ALGORITHM

(u,u2.u2) - (u.v.w

POISSON'S EQUATION

FE O ARE RETAINE LAPLACE'S EQUAT