Funded by the European Union

Co-ordinated by ECMWF

ESCAPE

Hardware Specific Optimizations in ESCAPE

Peter Messmer
pmessmer@nvida.com

MeteoSwiss · DMI · RMI · NVIDIA · METEO FRANCE Toujours un temps d'avance · ICHEC Irish Centre for High-End Computing · Deutscher Wetterdienst Wetter und Klima aus einer Hand · Optalysys · Bull atos technologies · PSNC · Loughborough University

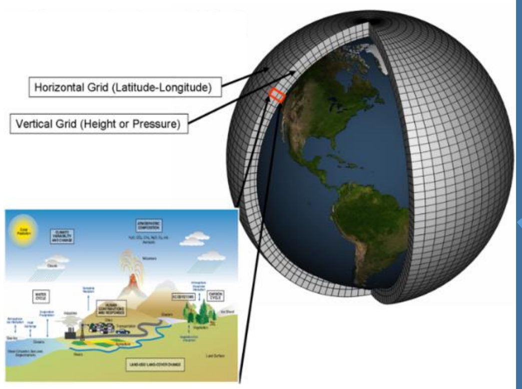
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 671627

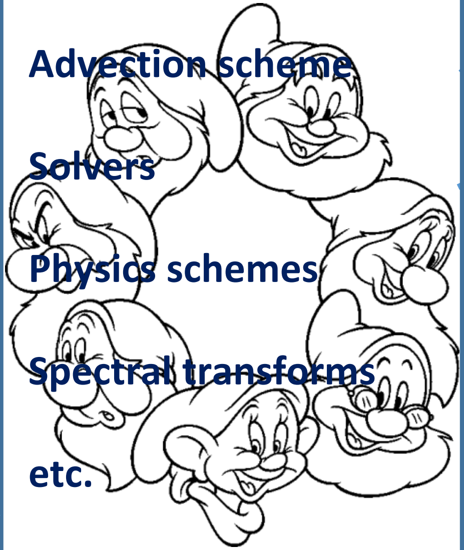
Funded by the European Union

ESCAPE

Energy efficient **SC**alable **A**lgorithms for weather **P**rediction at **E**xascale

Disassemble global … NWP model → **Extract, redesign… key components** → **Optimize for energy… efficiency on new hardware** → **… Reassemble global NWP model**





Advection scheme

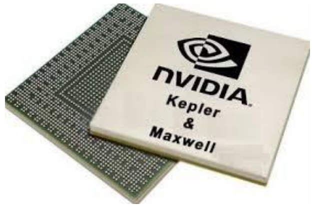
Solvers

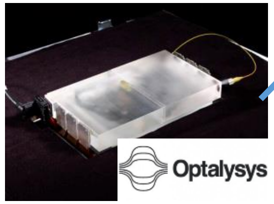
Physics schemes

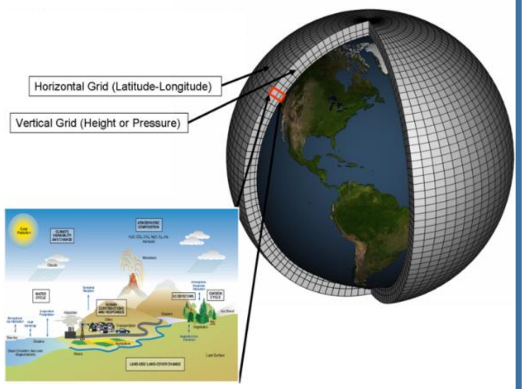
Spectral transforms

etc.

(memory, communication, compute intensity)



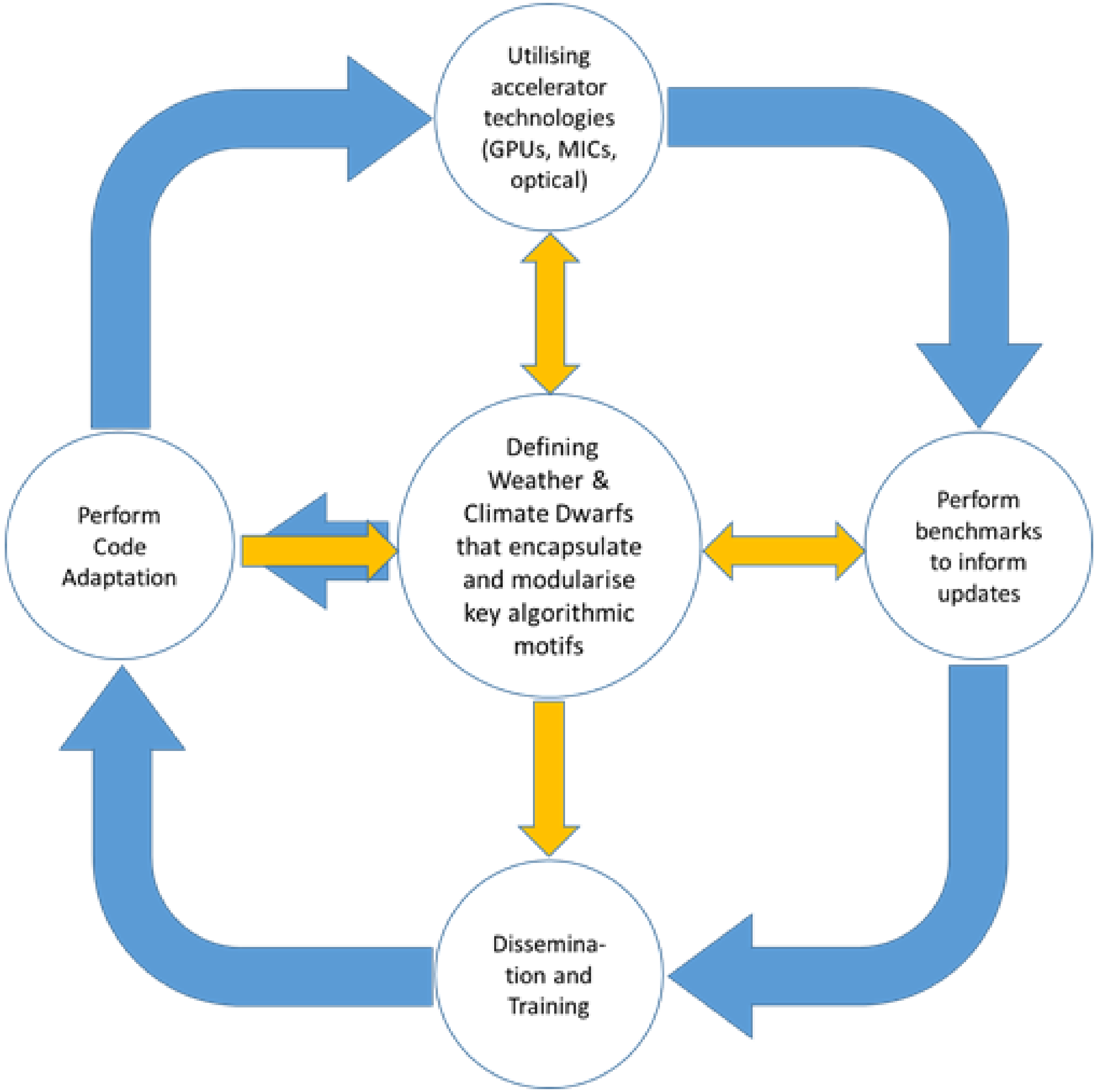




ECMWF · MeteoSwiss · DMI · RMI · NVIDIA · METEO FRANCE Toujours un temps d'avance · ICHEC · Deutscher Wetterdienst Wetter und Klima aus einer Hand DWD · Optalysys · Bull atos technologies · PSNC · Loughborough University

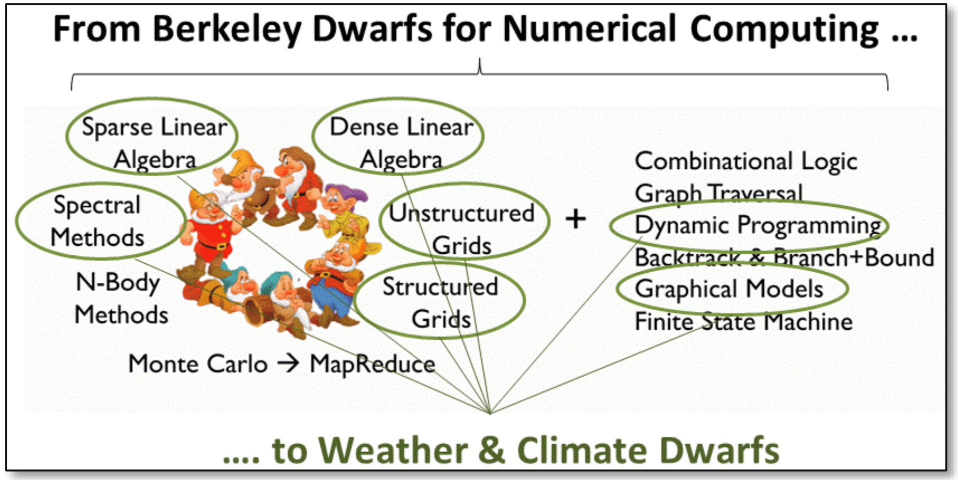
ESCAPE work flow

Funded by the European Union



ESCAPE promotes NWP & Climate dwarfs

Funded by the European Union



A dwarf encapsulates a **relevant characteristic or required functionality** of a weather/climate prediction model presented as **runnable** and **verifiable mini-application**

Dwarfs:

- Spectral transforms (FT/LT and bi-FT) …………. very memory and communication bandwidth intensive, possibly limited scalability
- 2 & 3-dimensional elliptic solver ………………... compute and communication latency intensive, possibly limited scalability
- Semi-Lagrangian advection ……………………... communication intensive, possibly limited scalability
- Flux-form finite-volume advection ………………. local communication, latency intensive, limited scalability
- Cloud physics parameterization …………………. expensive computations, independent vertical columns, scalable
- Radiation parameterization ………………………. expensive computations, spatial, temporal and wavenumber space, scalable
- More to come …

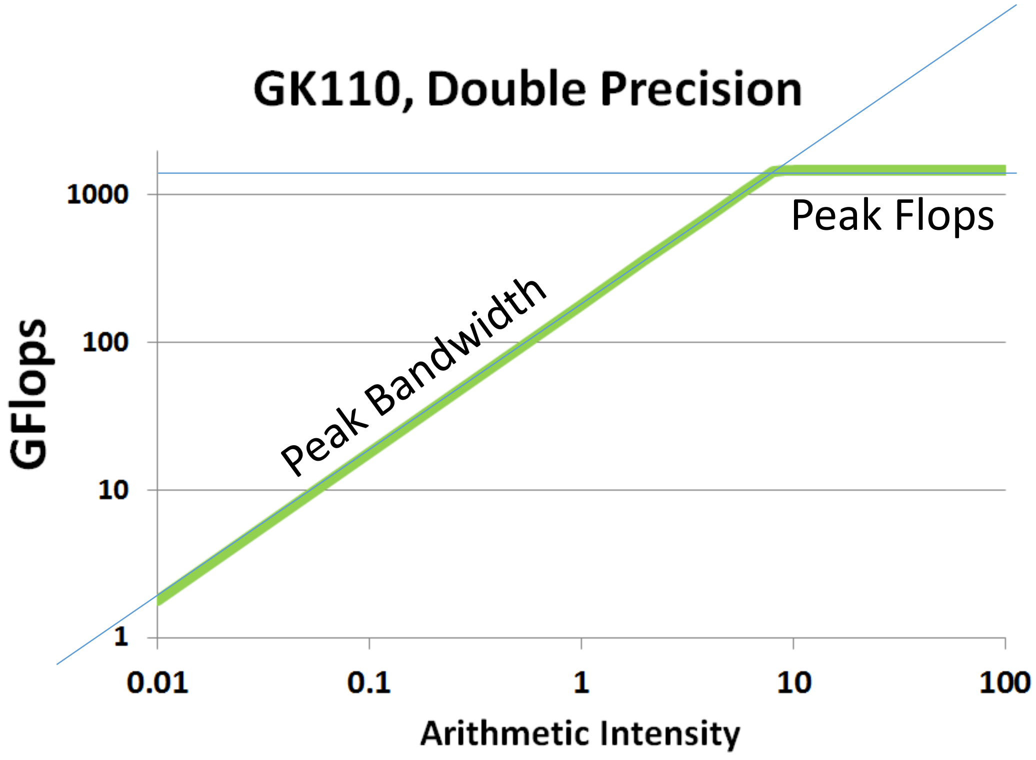
→ Dwarf design, accelerator adaptation, profiling, co-design, roofline modelling, …

ECMWF MeteoSwiss DMI RMI NVIDIA METEO FRANCE Toujours un temps d'avance ICHEC Deutscher Wetterdienst Wetter und Klima aus einer Hand Optalysys Bull PSNC Loughborough University

Performance models: Avoid fishing in the dark

Funded by the European Union

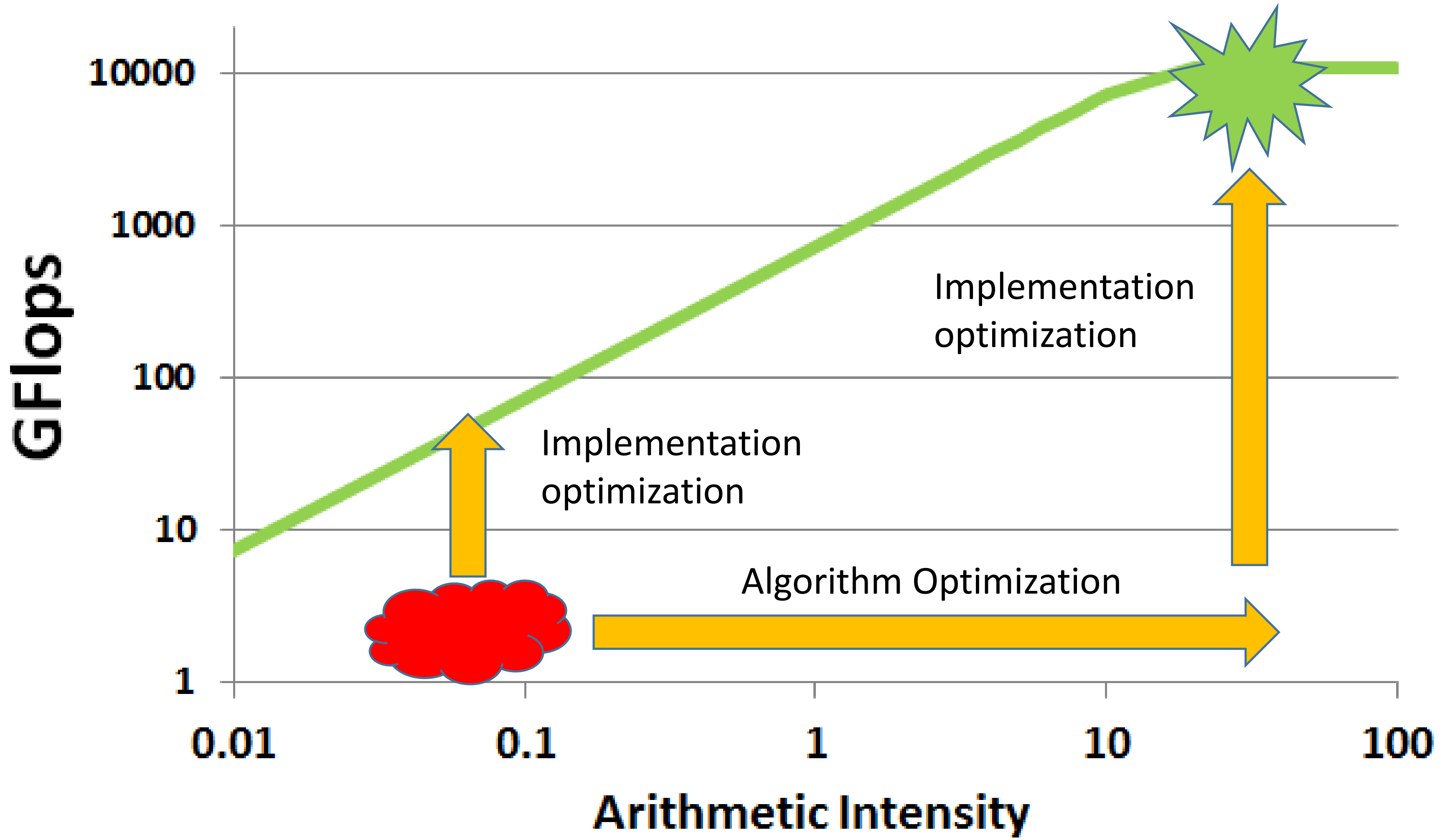
- **Low-level optimization can be hugely time consuming**
  - Detailed hardware dependency
  - Depending on software stack
- **Optimization potential estimated by roofline model**
- ***UNDERSTAND* performance limiters**
  - **Design optimizations**
  - **Model to project achievable performance**



Arithmetic Intensity: #FLOPS/#Bytes

Optimization: Algorithm, Implementation

Funded by the European Union



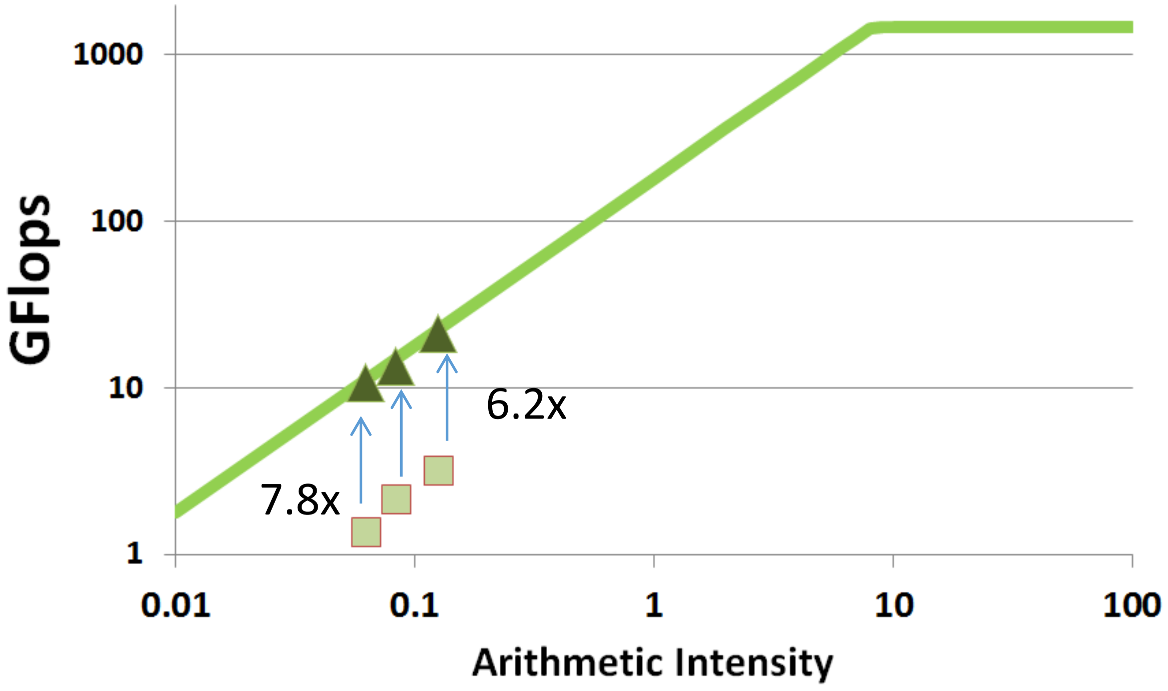
Dwarf #2 GPU Optimization: Loop Reordering

Funded by the European Union

- **Optimization of data access pattern**
- **Leads to optimal bandwidth utilization**

After:

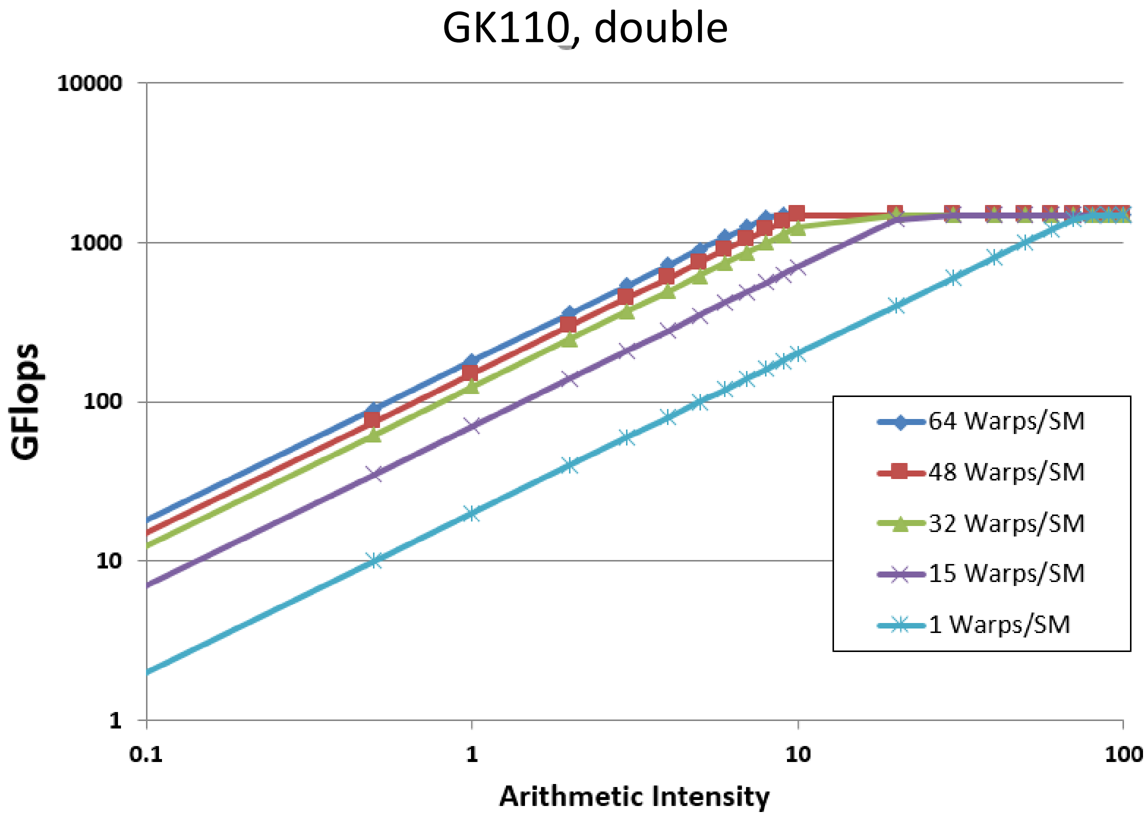
```
!$acc parallel loop collapse(2)
      do j=1,ny
        do i=1,nx
          x (i,j,l+1) = x (i,j,l+1) + del(ll) * x (i,j,ll)
          ax(i,j,l+1) = ax(i,j,l+1) + del(ll) * ax(i,j,ll)
        enddo
      enddo
!$acc end parallel
```



Roofline Model: Knowing your limits

Funded by the European Union

- "Warp" – instruction for group of 32 threads
- GK110: Up to 64 warps at various stages of execution per SM
- Achievable bandwidth depends on number of warps in flight
  - Instruction mix
  - Number of warps per SM ("occupancy")
  - Number of warps in total ("utilization")

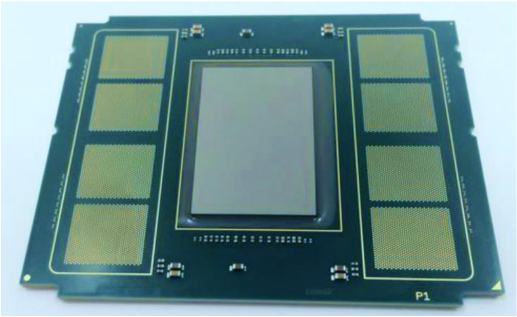


Hybrid Computing Partners

Funded by the European Union

Bull atos technologies

Xeon/Xeon Phi

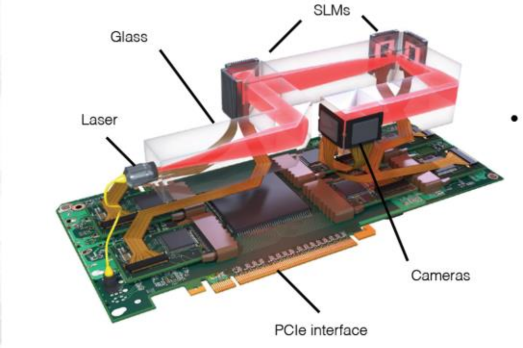


NVIDIA

GPU

Optalysys

Optical Processor

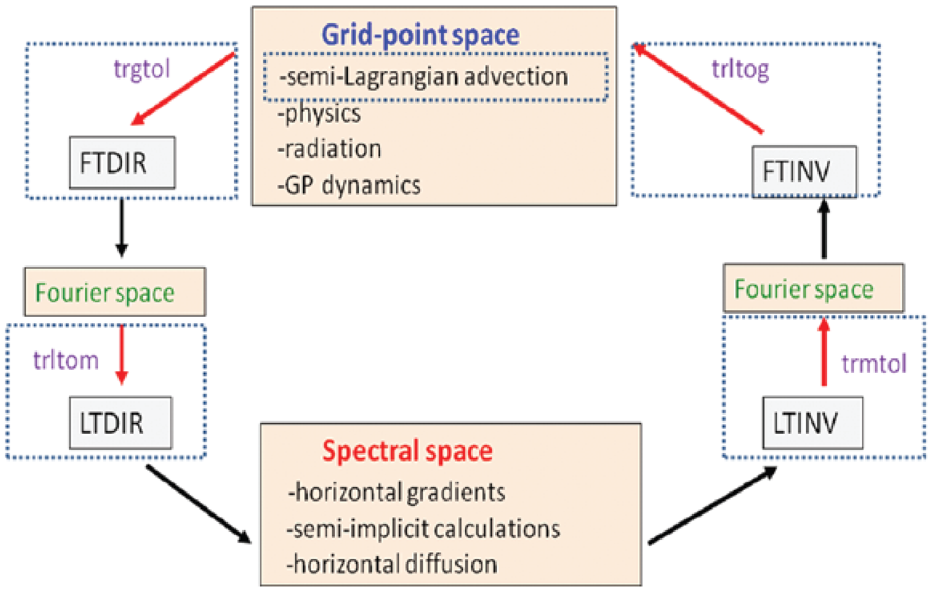


Performance Model PSNC **Energy Efficiency**

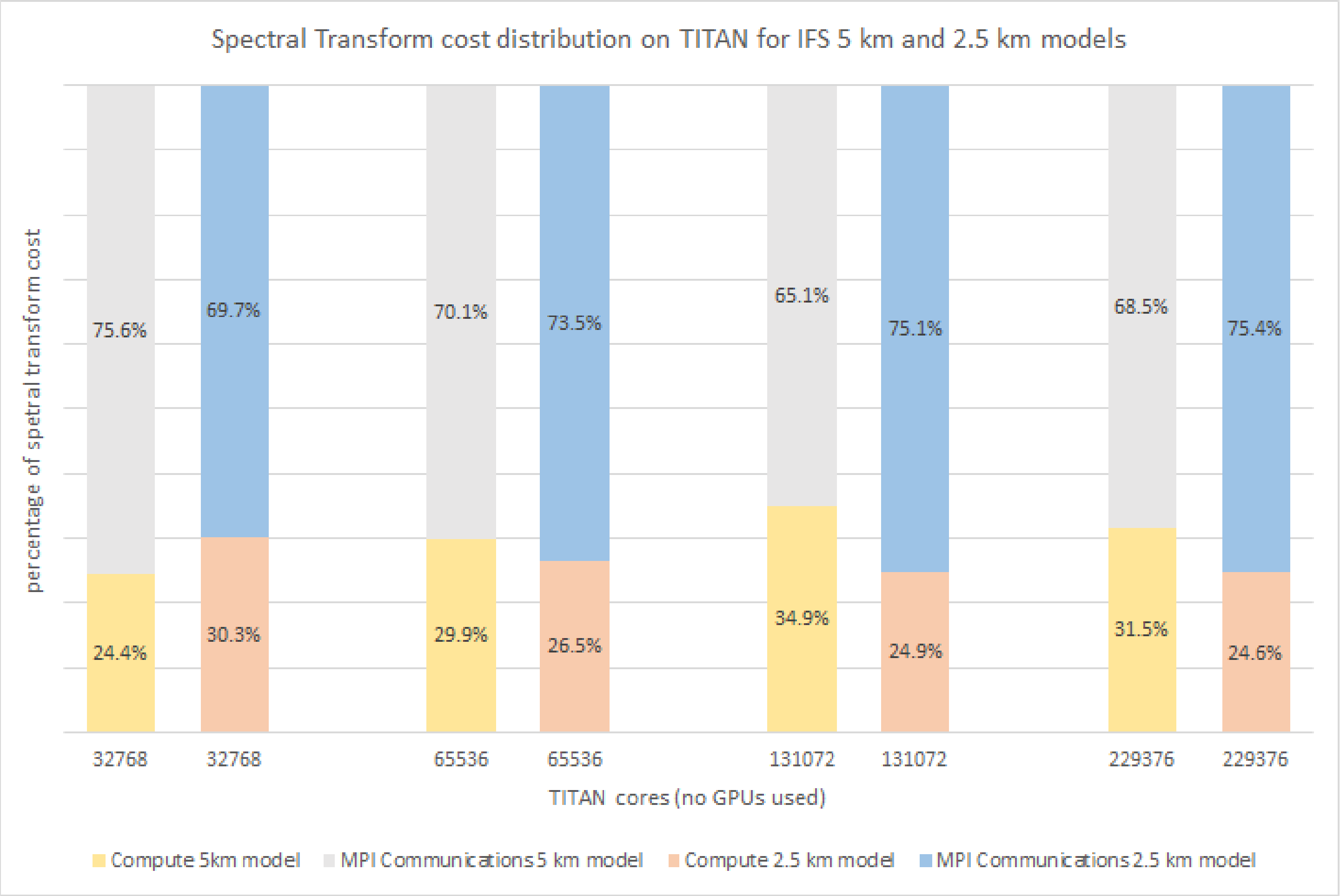
Example Dwarf #1: Spectral transforms

Funded by the European Union

Once per time step:



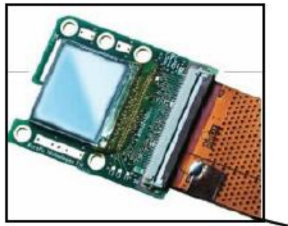
Communication vs computation cost:

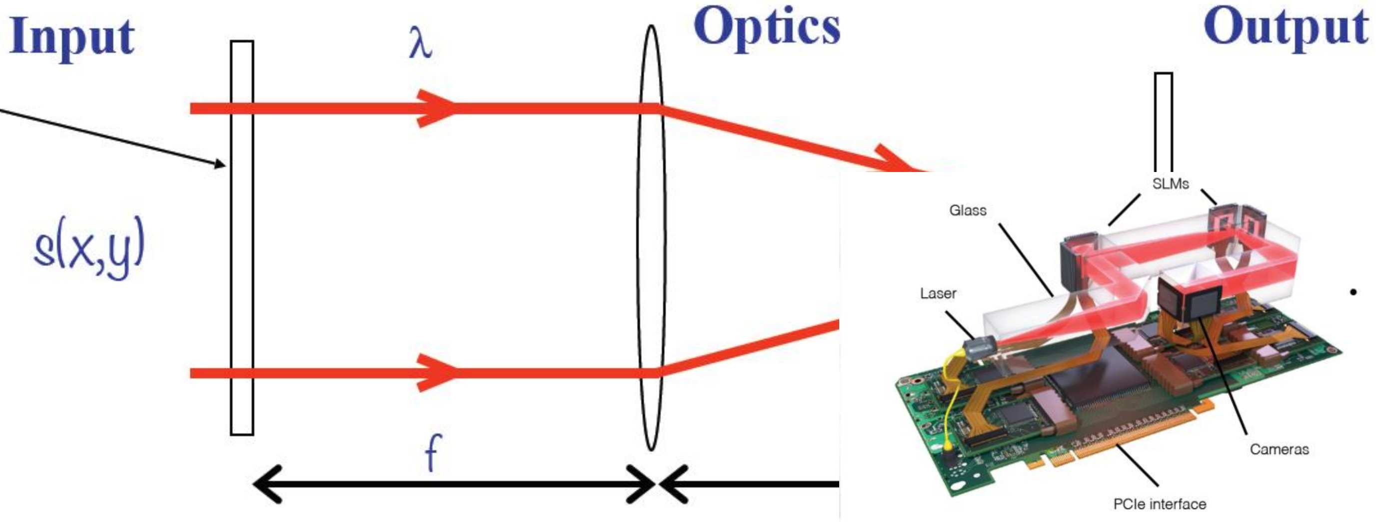


[Courtesy George Mozdzynski]

Optical Processing and Liquid Crystal SLMs

Funded by the European Union





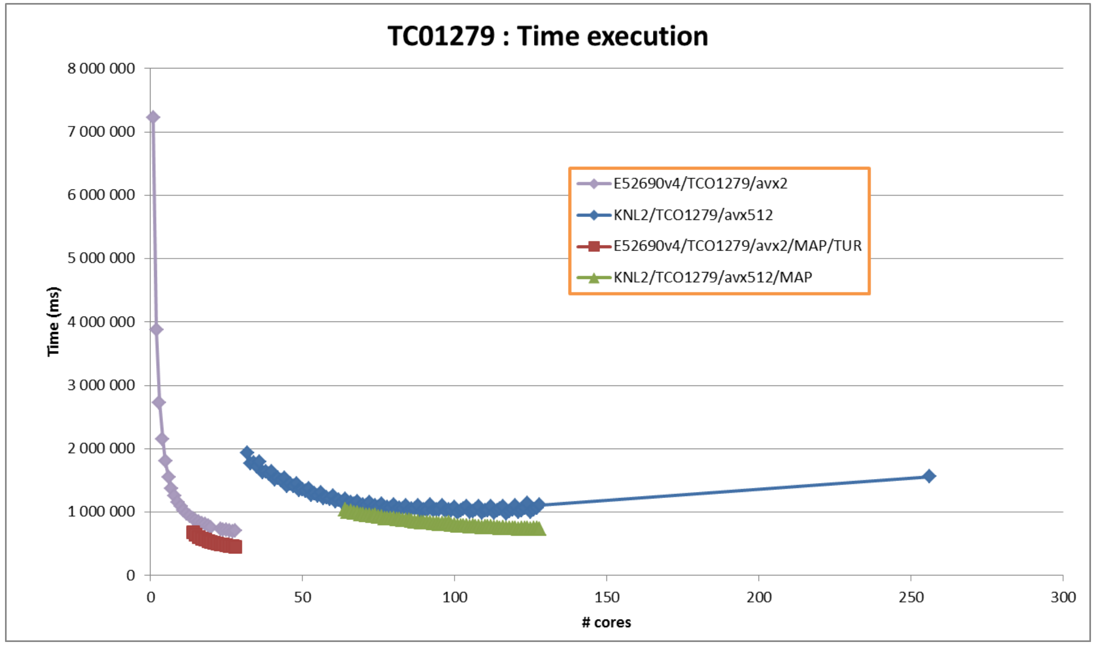
- Spatial Light Modulators (SLMs) used to enter numerical data
- Typical power of <5W/SLM

$$S(u,v) = FT[s(x,y)] = \iint_{\pm\infty} s(x,y) \times \exp\left(-j2\pi\left[xu + yv\right]\right) dxdy$$

Spectral transform: Initial Timings on Xeon, KNL

Funded by the European Union

- Dwarf-D-spectralTransform-sphericalHarmonics (Atlas proto)
  - Test case: TCo1279 (9 km), same as the HRES operational model at ECMWF



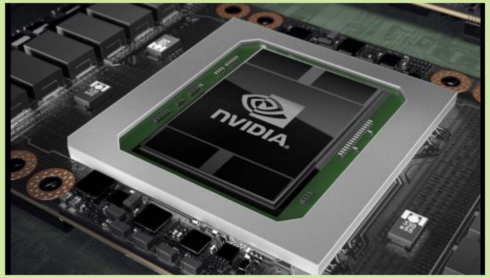
Intel Xeon E52690v4 (BDW), dwarf compiled with AVX2 support (light purple), with turbo enabled and mmap disabled (red).

Intel Xeon Phi 7210 (KNL) in cache mode quadrant, dwarf compiled with AVX512 support with turbo enabled (blue), with mmap disabled (green).

INTRODUCING TESLA P100

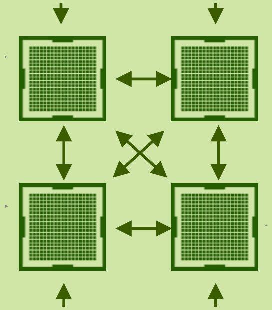
New GPU Architecture to Enable the World's Fastest Compute Node

Pascal Architecture



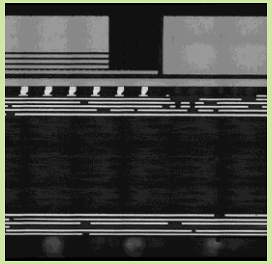
Highest Compute Performance

NVLink



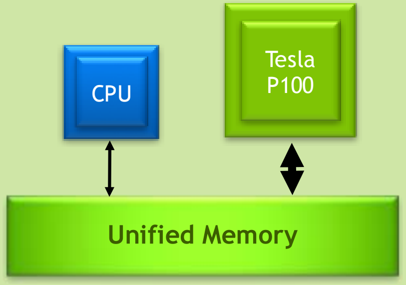
GPU Interconnect for Maximum Scalability

CoWoS HBM2

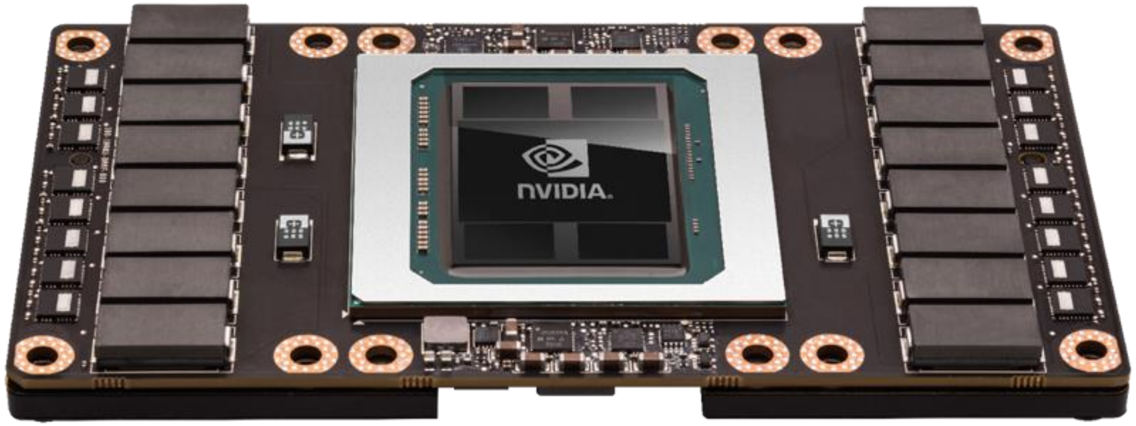


Unifying Compute & Memory in Single Package

Page Migration Engine



Simple Parallel Programming with Virtually Unlimited Memory



Tesla P100 GPU: GP100

56 SMs

3584 CUDA Cores

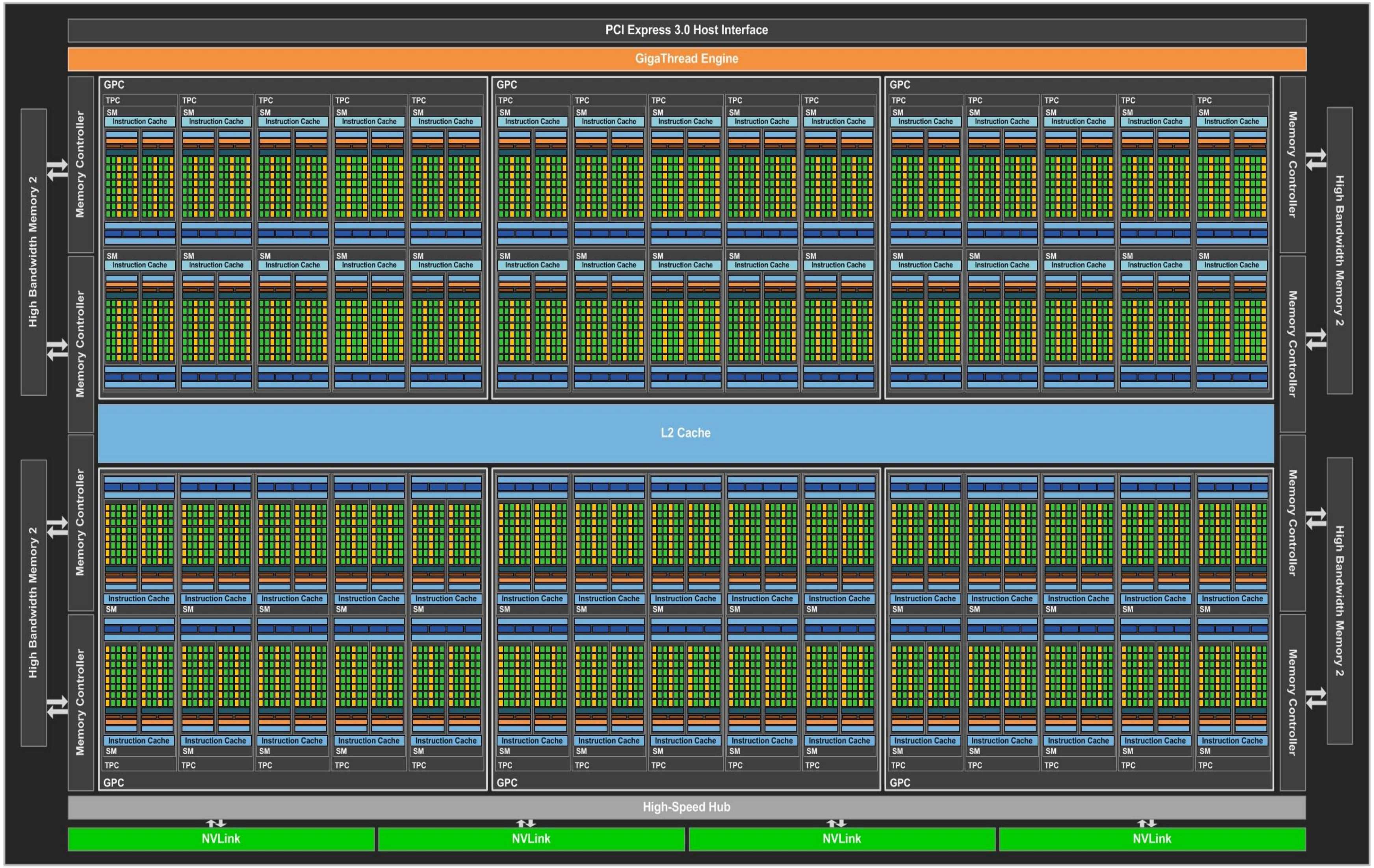
5.3 TF Double Precision

10.6 TF Single Precision

21.2 TF Half Precision

16 GB HBM2

720 GB/s Bandwidth



GPU Performance Comparison

	P100	M40	K40
Double Precision TFlop/s	5.3	0.2	1.4
Single Precision TFlop/s	10.6	7.0	4.3
Half Precision Tflop/s	21.2	NA	NA
Memory Bandwidth (GB/s)	720	288	288
Memory Size	16GB	12GB, 24GB	12GB

NVLink

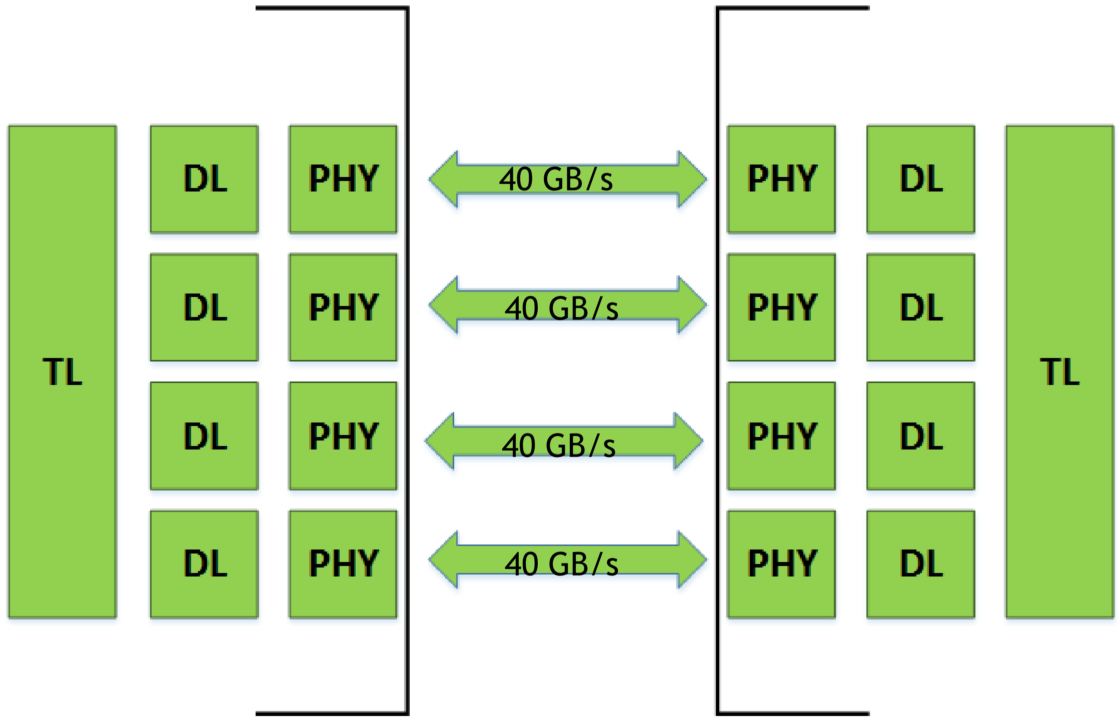
P100 supports 4 NVLinks

Up to 94% bandwidth efficiency

Supports read/writes/atomics to peer GPU

Supports read/write access to NVLink-enabled CPU

Links can be ganged for higher bandwidth



NVLink on Tesla P100

DGX-1: NVLink coupled GPU Cluster

Two fully connected quads, connected at corners

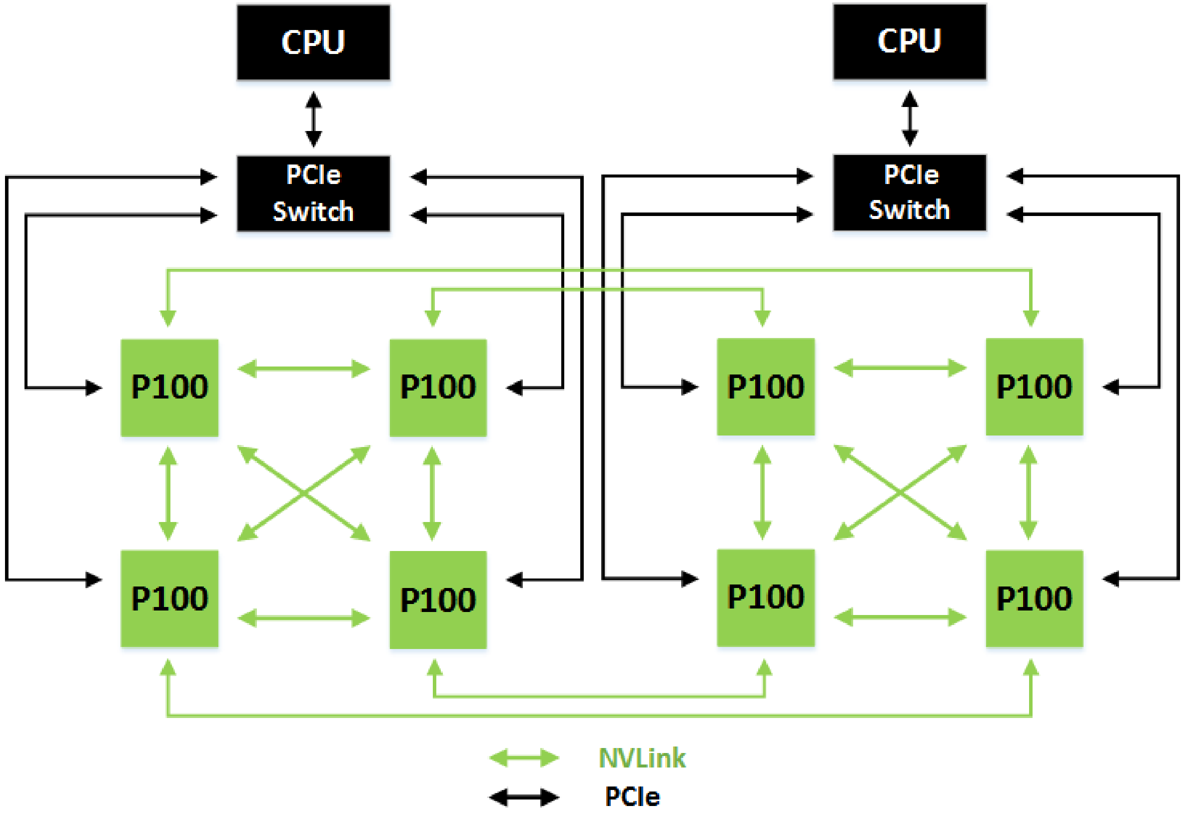
160GB/s per GPU bidirectional to Peers

Load/store access to Peer Memory

Full atomics to Peer GPUs

High speed copy engines for bulk data copy

PCIe to/from CPU



NVLink to CPU

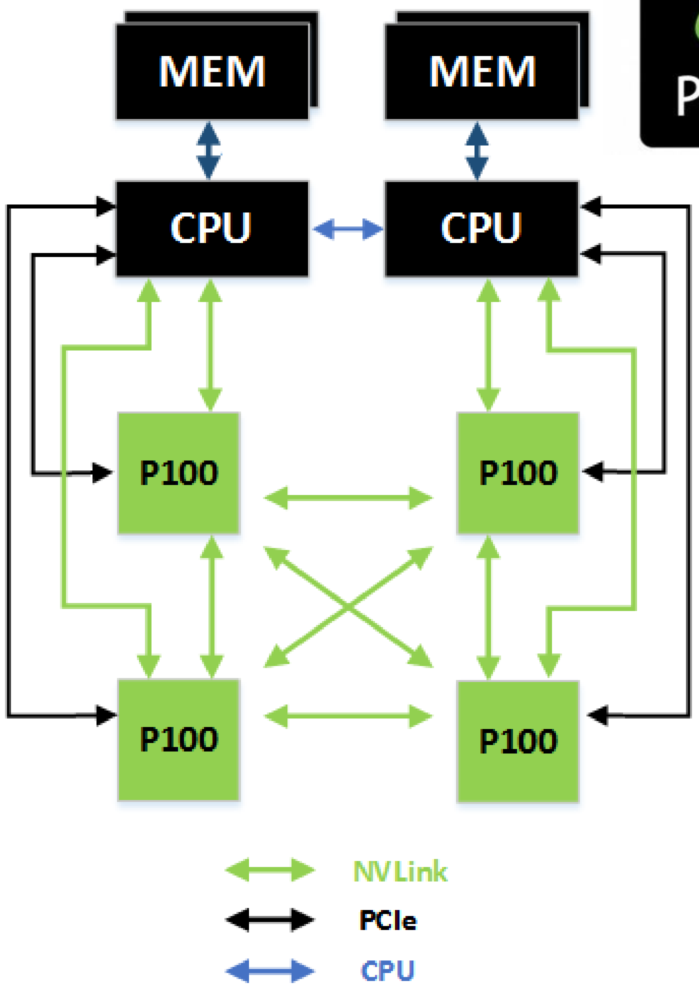
Fully connected quad

120 GB/s per GPU bidirectional for peer traffic

40 GB/s per GPU bidirectional to CPU

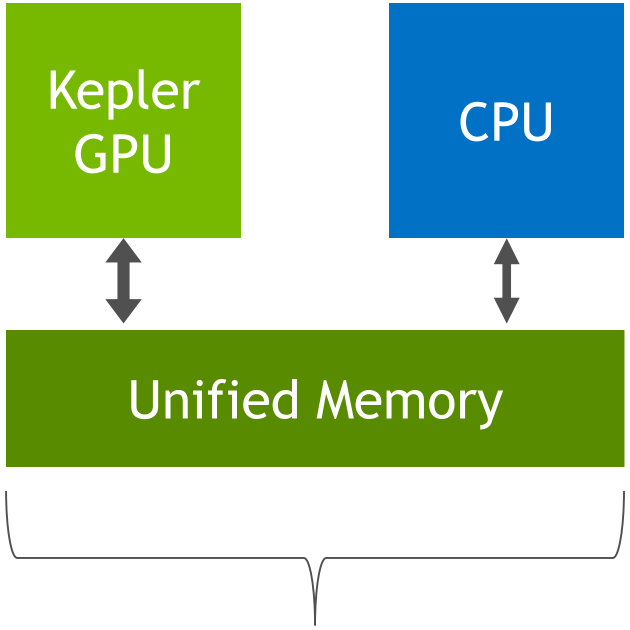
Direct Load/store access to CPU Memory

High Speed Copy Engines for bulk data movement



Kepler/Maxwell Unified Memory

CUDA 6+



Allocate Up To
GPU Memory Size

Simpler Programming & Memory Model

- Single allocation, single pointer, accessible anywhere
- Eliminate need for *explicit copy*
- Greatly simplifies code porting

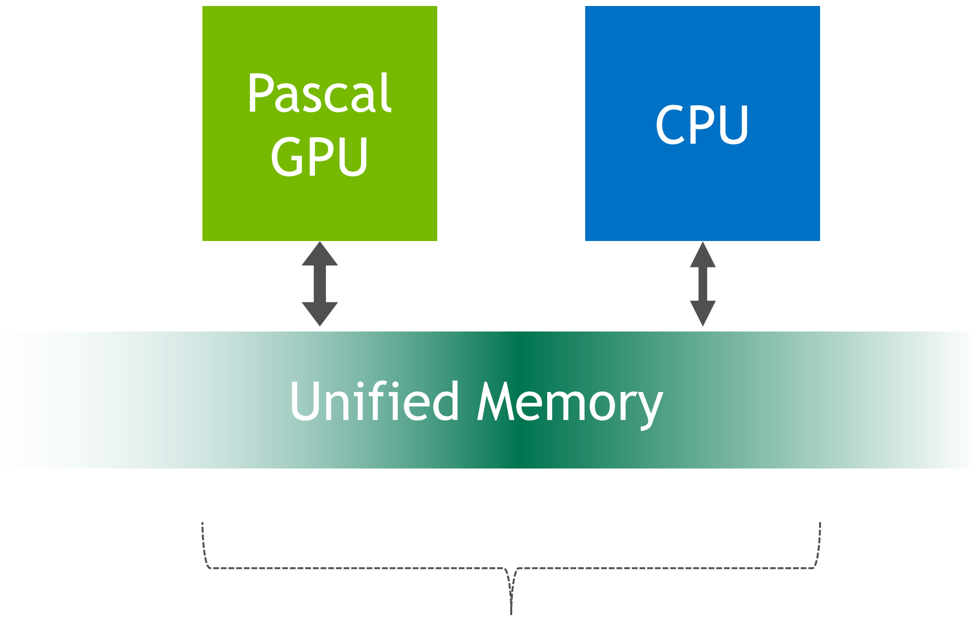
Performance Through Data Locality

- Migrate data to accessing processor
- Guarantee global coherency
- Still allows explicit hand tuning

Pascal Unified Memory

Large datasets, simple programming, High Performance

CUDA 8



Allocate Beyond GPU Memory Size

Enable Large Data Models

- Oversubscribe GPU memory
- Allocate up to system memory size

Tune Unified Memory Performance

- Usage hints via cudaMemAdvise API
- Explicit prefetching API

Simpler Data Access

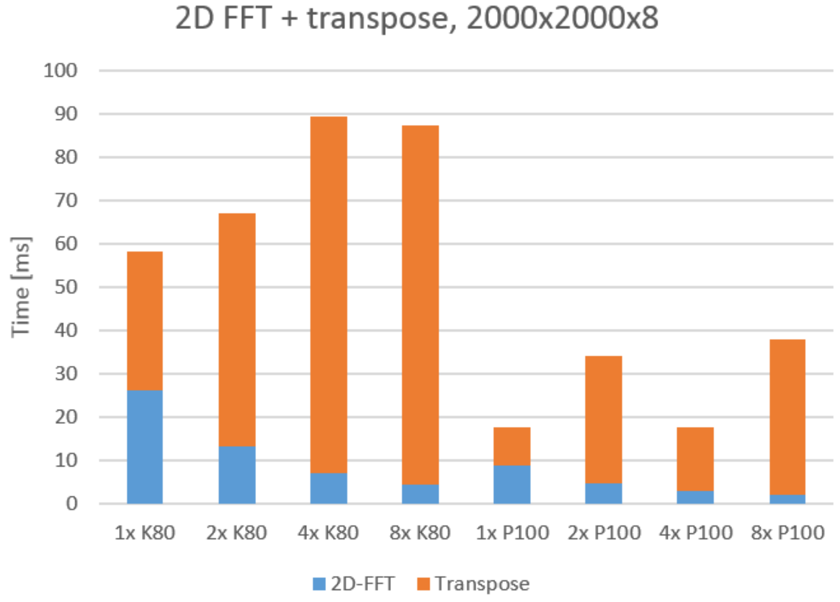
- CPU/GPU Data coherence
- Unified memory atomic operations

Multi-GPU Spectral Transform Proxy

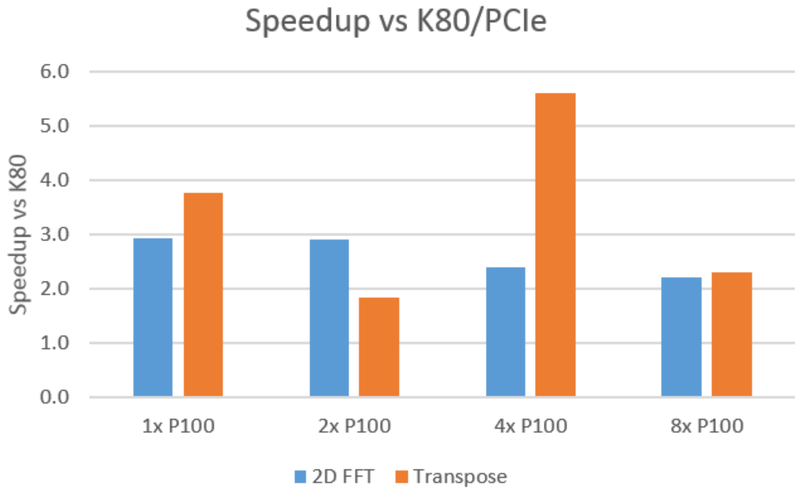
Funded by the European Union

- **Proxy for Spectral transform**
  - Batched 2D FFT
  - Batched 1D FFT, followed by GEMM
- **Typically followed by transpose**

  (collocated planes <-> collocated columns)
- **Transpose: 5x speedup for NVLink vs PCIe**
- **2D FFT: ~ 3x speedup vs K80**
- **Analysis and further optimization ongoing**



Preliminary



Preliminary

Summary

Funded by the European Union

- **Platform specific optimizations of dwarfs key part of ESCAPE project**
  - Explore speed of light on given platform
  - Performance projections to future systems

- **Broad spectrum of architectures**
  - Optalysys Optical Processor, Xeon, Xeon Phi, GPU

- **Key dwarfs identified**
  - Spectral transform, bi-FFT

- **Investigations of latest generation GPU architectures**
  - NVLink, HBM2, FP16
  - Initial multi-GPU investigations under way