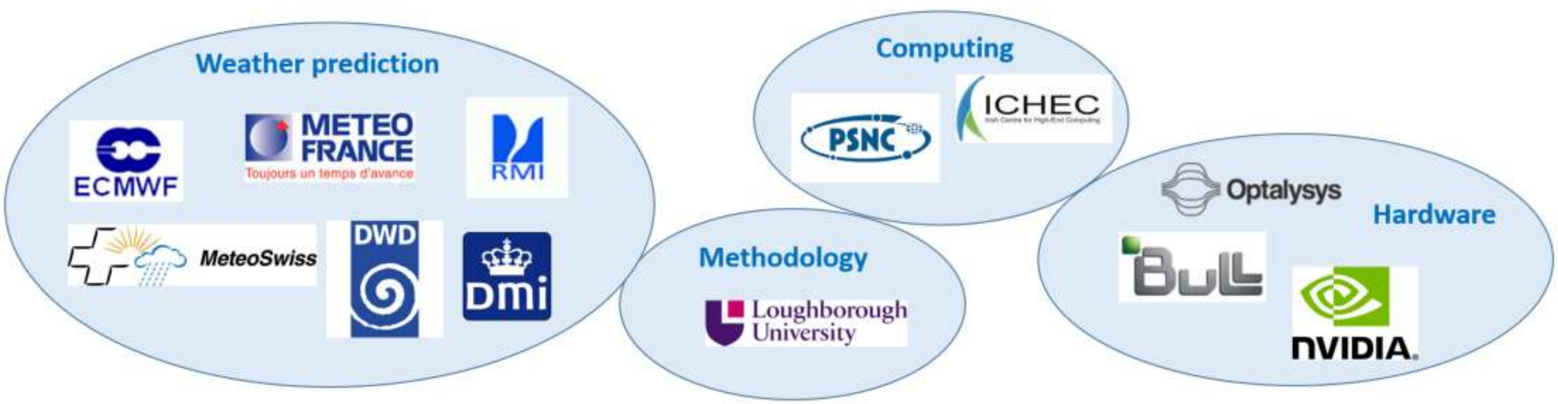
ACCELERATING WEATHER PREDICTION WITH NVIDIA GPUS

Alan Gray, Developer Technology Engineer, NVIDIA

ECMWF 18th Workshop on high performance computing in meteorology, 28th September 2018

ESCAPE

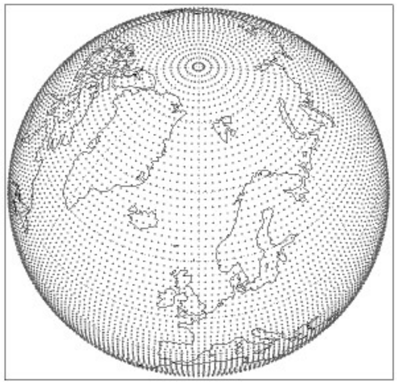
Energy-efficient **Sc**alable **A**lgorithms for Weather **P**rediction at **E**xascale



- NVIDIA's role is to take existing GPU-enabled codes and optimize.

ESCAPE DWARVES

Spherical Harmonics (SH) Dwarf

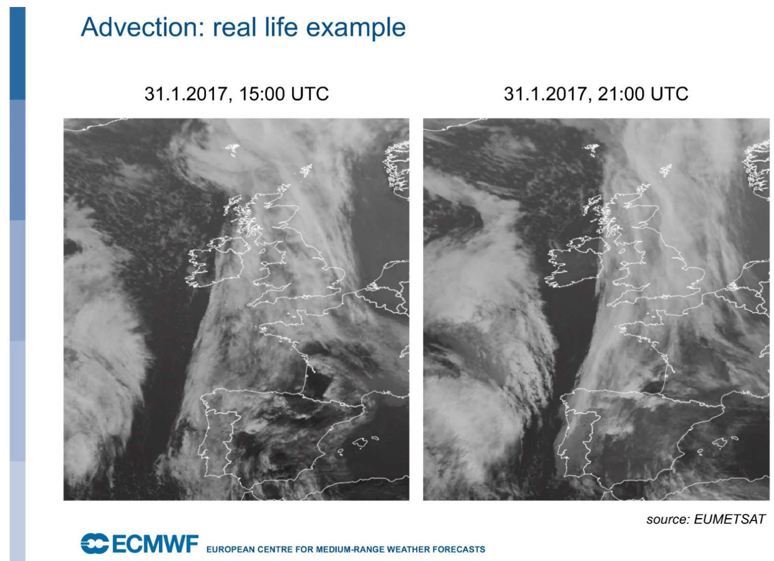


- ECMWF’s Integrated Forecasting System (IFS) is a global prediction system: entire earth’s atmosphere is represented as a spherical grid.
- Info in “grid-point” space can be equivalently represented in “spectral” space, i.e. in terms of the frequencies of the fluctuating waves, which is more suited to some calculations.
- IFS therefore repeatedly transforms between these representations, **Fourier transforms** (FFTs) in longitude and **Legendre transforms** (DGEMMs) in latitude, with AlltoAll data movement in-between.
- This dwarf represents the spectral transforms from IFS.
- NB. Number of points varies (e.g. most round equator, fewest at poles). Additionally, there exist multiple altitude “levels”, in third dimension away from surface of earth, each with 3 “fields”.

ESCAPE DWARVES

MPDATA Dwarf

- Advection: horizontal transport
- Uses unstructured grid with nearest-neighbour stencils
- MPDATA scheme already used within COSMO-EULAG (PSNC), and of interest to ECMWF for future developments



- Both SH and MPDATA Dwarves Fortran+OpenACC+MPI. SH also has interfacing to CUDA libraries.
- **Many of the optimizations I will present are transferable to other applications/languages etc.**

SINGLE GPU OPTIMIZATION

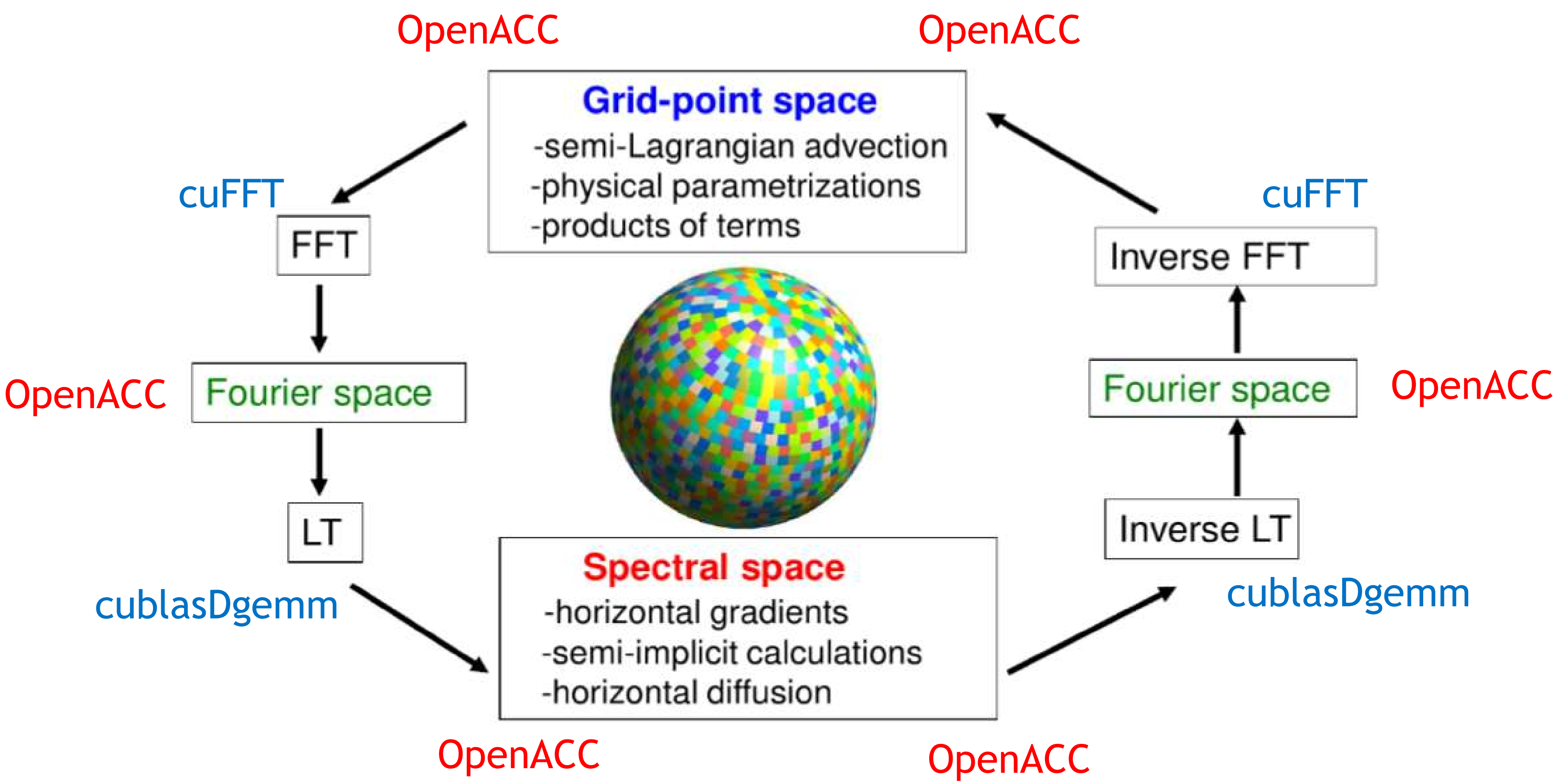
Exposing Parallelism: Original implementations had naïve mapping of loops to the GPU, and the resulting decompositions did not map well. We have restructured to tightly nested loops, and used “collapse” OpenACC clause to allow compiler to map all inherent parallelism to hardware in an efficient manner.

Optimizing data management such that the fields stay resident on the GPU for the whole timestep loop: all allocations/frees have been moved outside the timestep loop with temporary work arrays being re-used, and all host/device data transfer has been minimized.

Memory Coalescing: Restructuring of array layouts to ensure memory coalescing. Sometimes transposes necessary: use OpenACC “tile” clause or push into BLAS library where possible.

For full details see GTC18 recording at http://on-demand-gtc.gputechconf.com/gtc-quicklink/2JS6yr

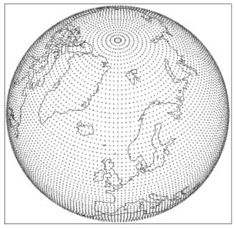
INTEROPERABILITY AND LIBRARIES: SH DWARF



Base language Fortran, MPI for multi-GPU communications.

BLAS/FFT LIBRARY CALLS IN SH DWARF

- At each timestep, SH dwarf performs transforms using Matrix Multiplications and FFTs.
- Multiple operations - one for each:
  - Field (associated with vertical levels)
  - Longitude (Matmult) / Latitude (FFT)



- Can batch over fields, since sizes are the same. But different longitudes/latitudes have different sizes: not supported by batched versions of cublasDgemm/cuFFT.
  - So, originally we had many small calls: low parallelism exposure and launch latency sensitivity.
- For DGEMM, we pad with zeros up to largest size and batch over longitudes as well as fields: single call to library; extra operations do not contribute to result.
- But FFT does not allow padding in the same way. Worked around launch latency problem by removing sync after each call: allows launch latency to be hidden behind execution.
  - As will be seen, however, this is the only part of the dwarf which remains suboptimal. Future: batched FFT with differing sizes should improve performance.

MPDATA OPTIMIZATION: P100

Before:

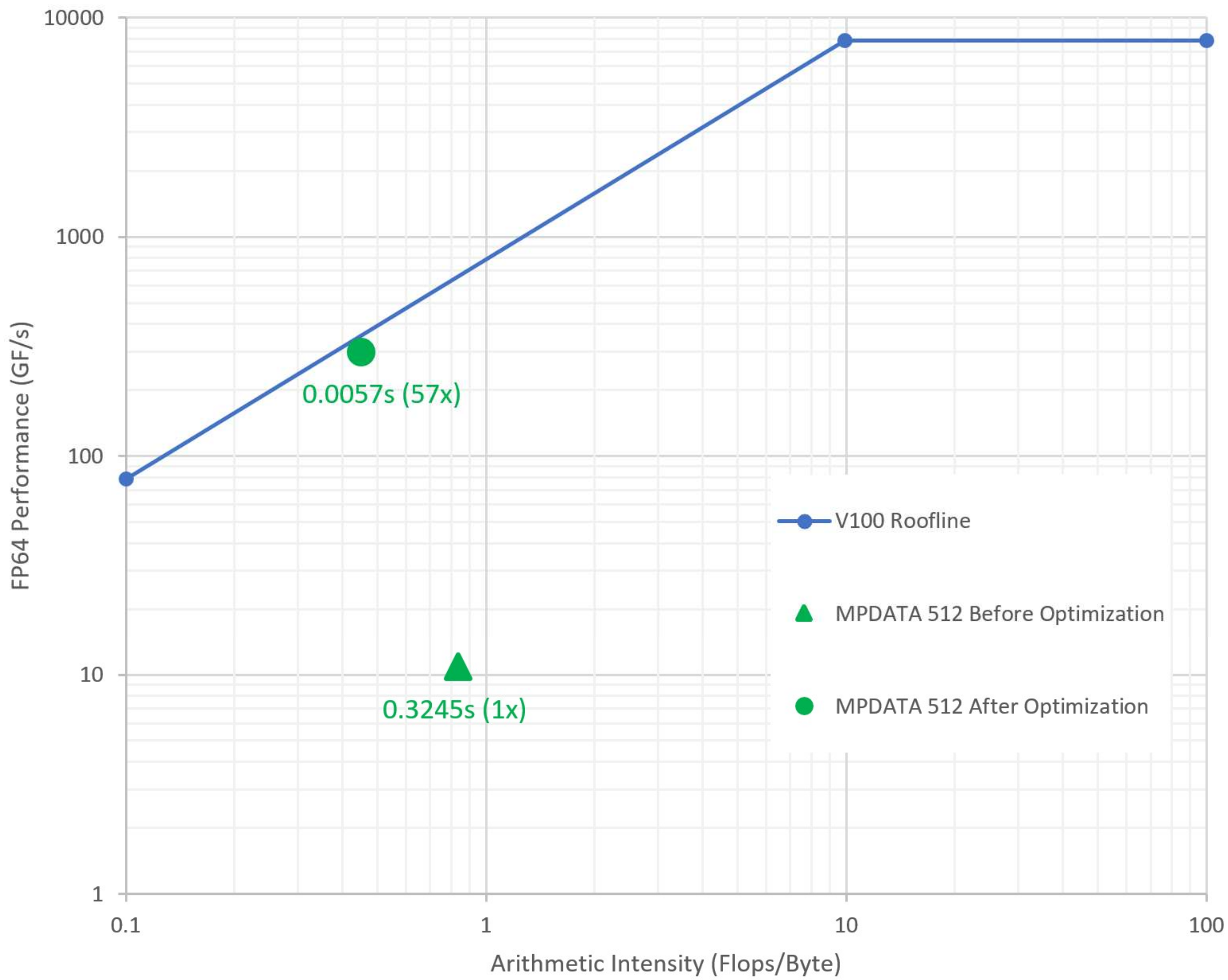
After:

OPTIMIZED MPDATA: P100 VS V100

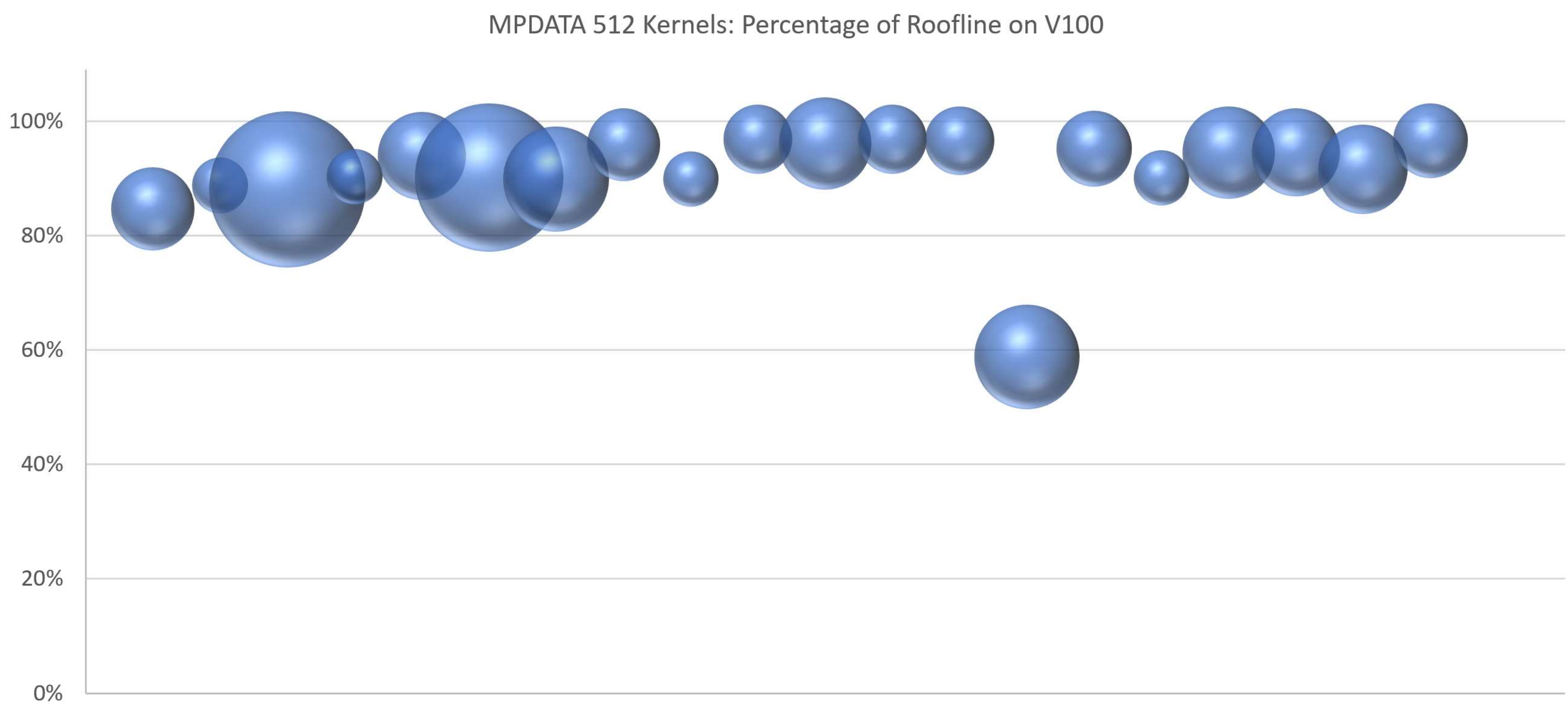
P100

V100

ESCAPE DWARF V100 PERFORMANCE

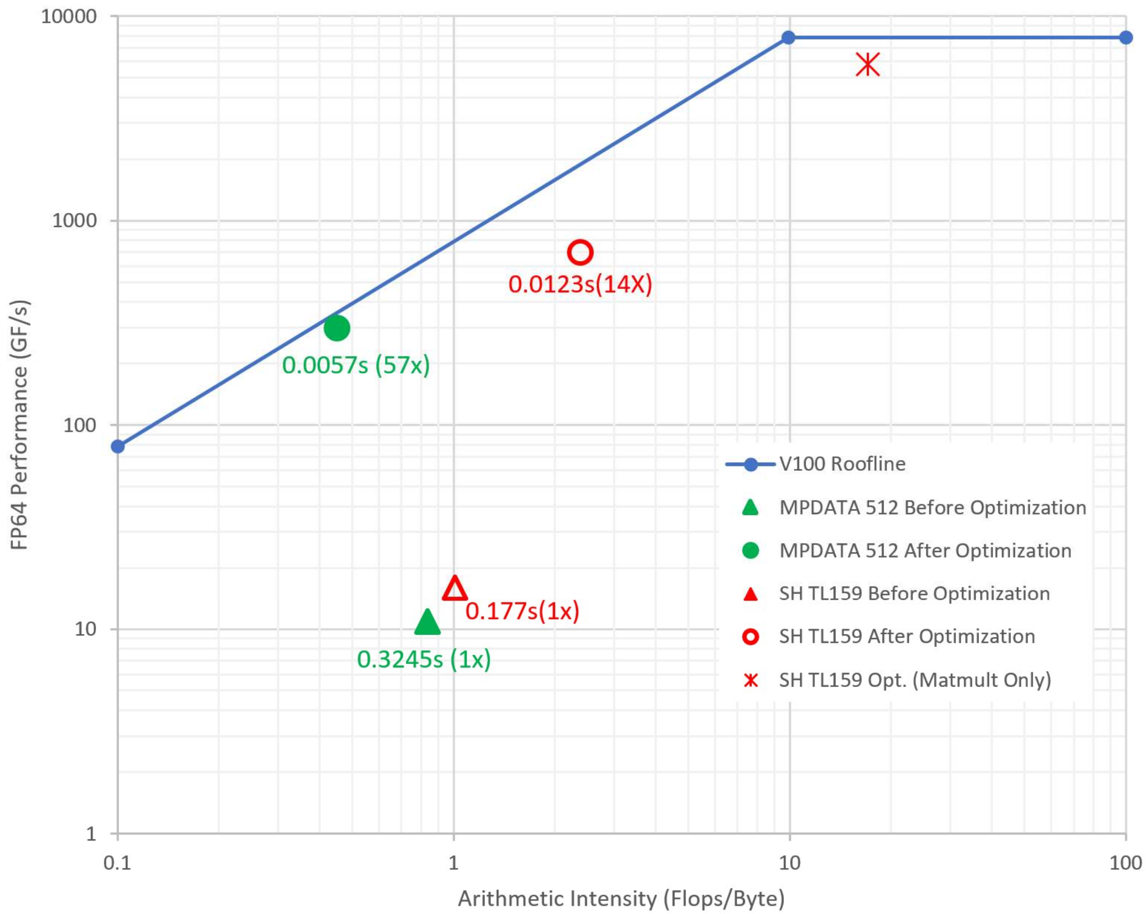


MPDATA KERNEL PERFORMANCE

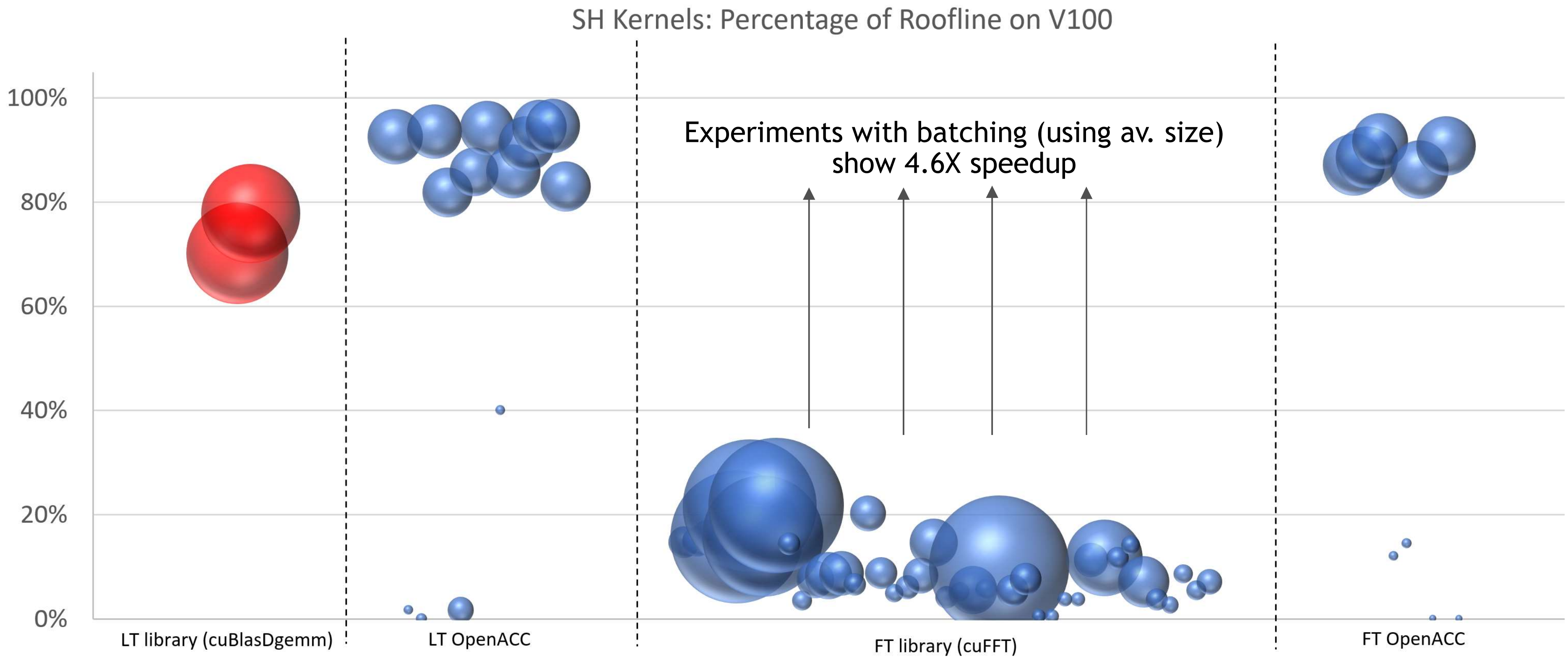


- 100% Roofline is STREAM benchmark throughput, since all kernels are memory bandwidth bound

ESCAPE DWARF V100 PERFORMANCE

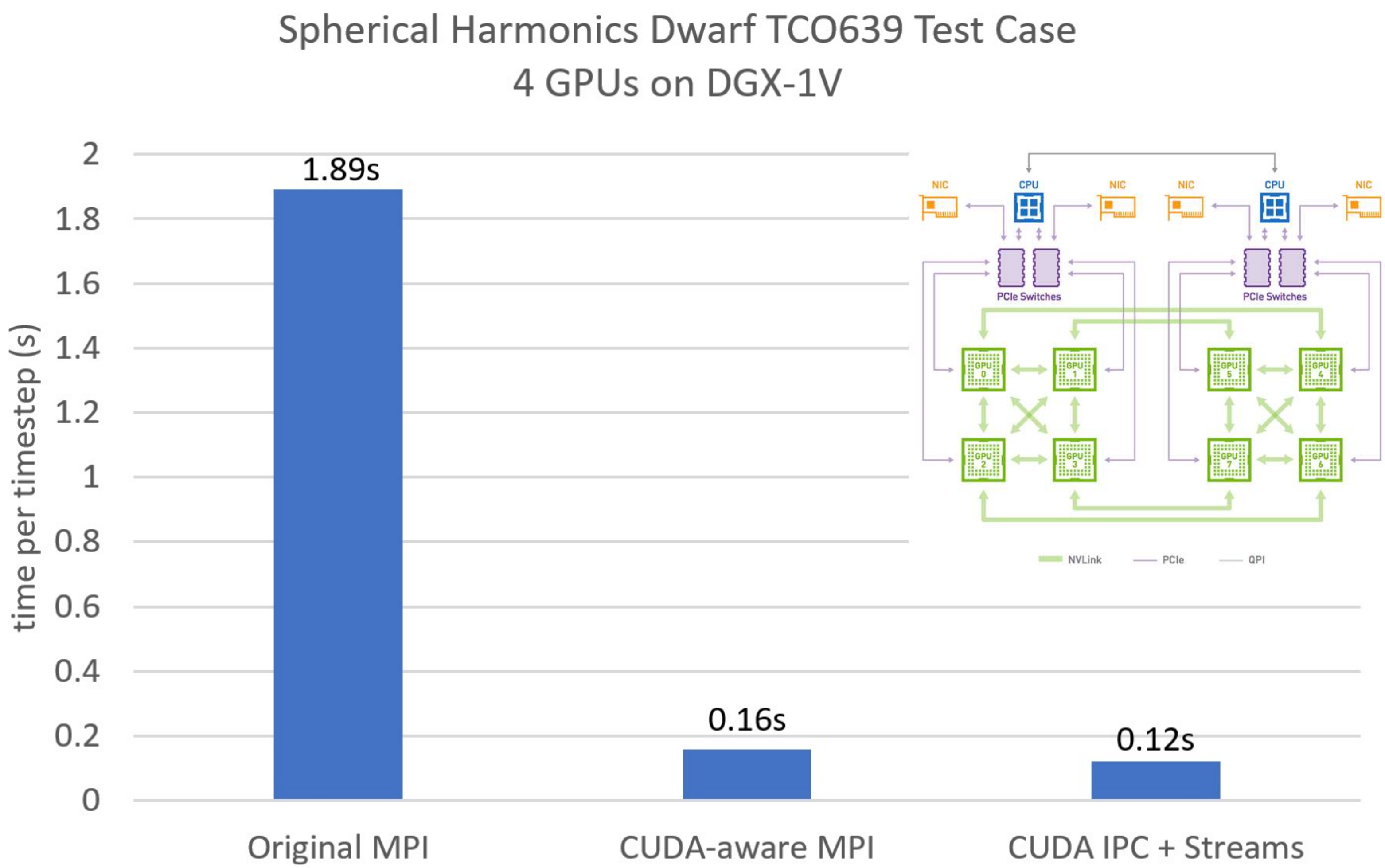


SH KERNEL PERFORMANCE

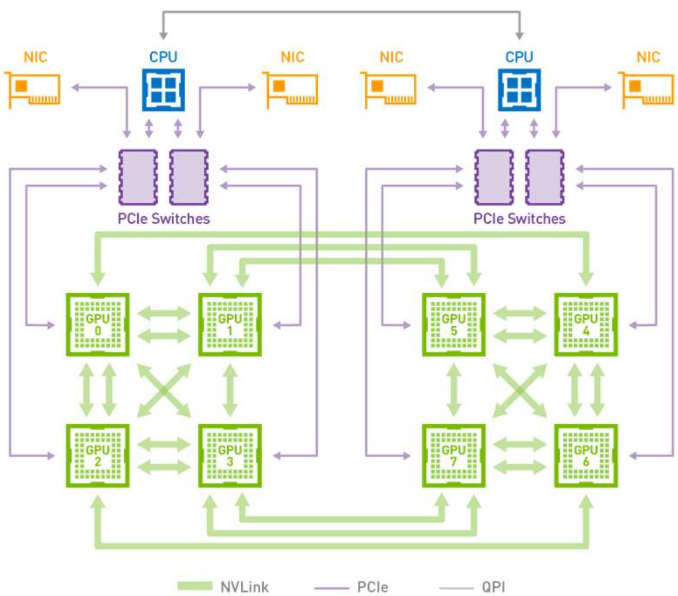


- 100% Roofline is peak DP Performance (compute bound kernels) or STREAM benchmark throughput (memory bandwidth bound kernels)

SH RESULTS ON 4 GPUS

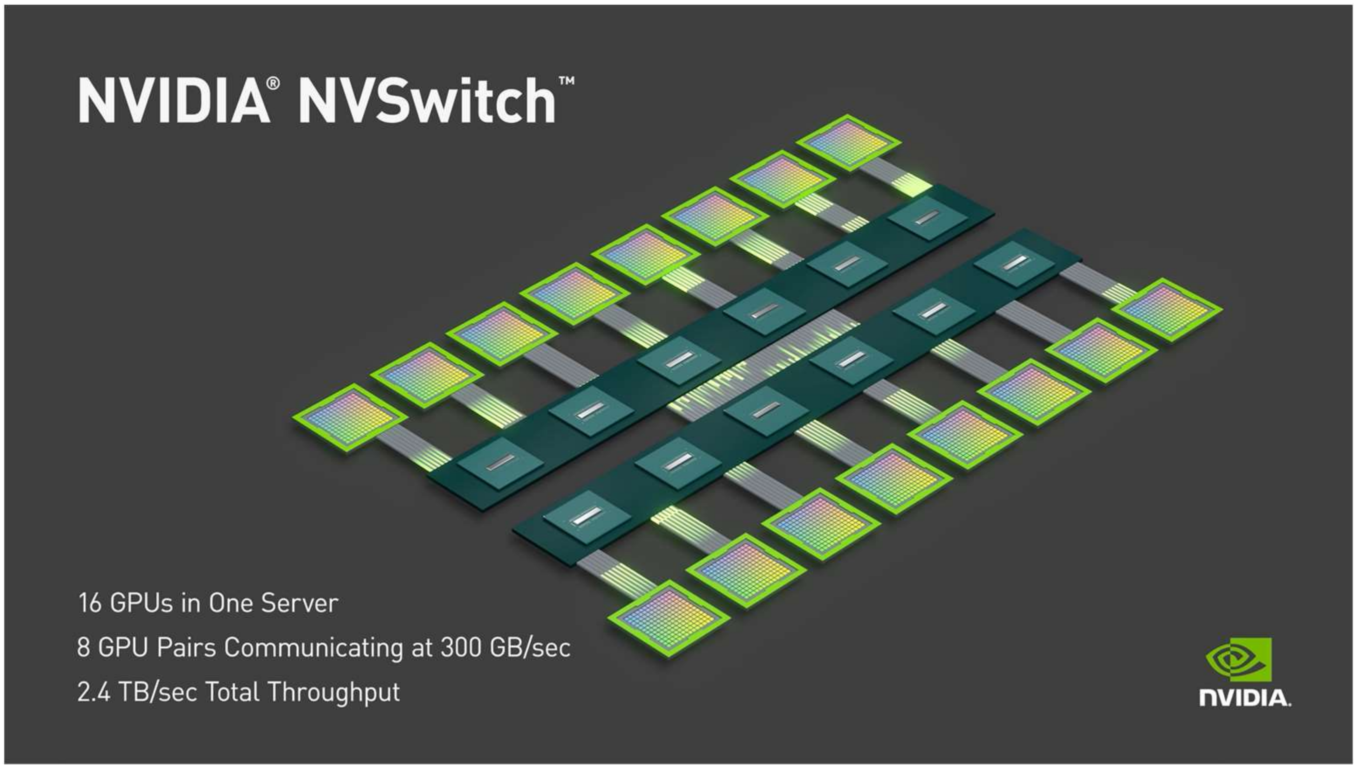


SPHERICAL HARMONICS: SCALING BEYOND 4 GPUS



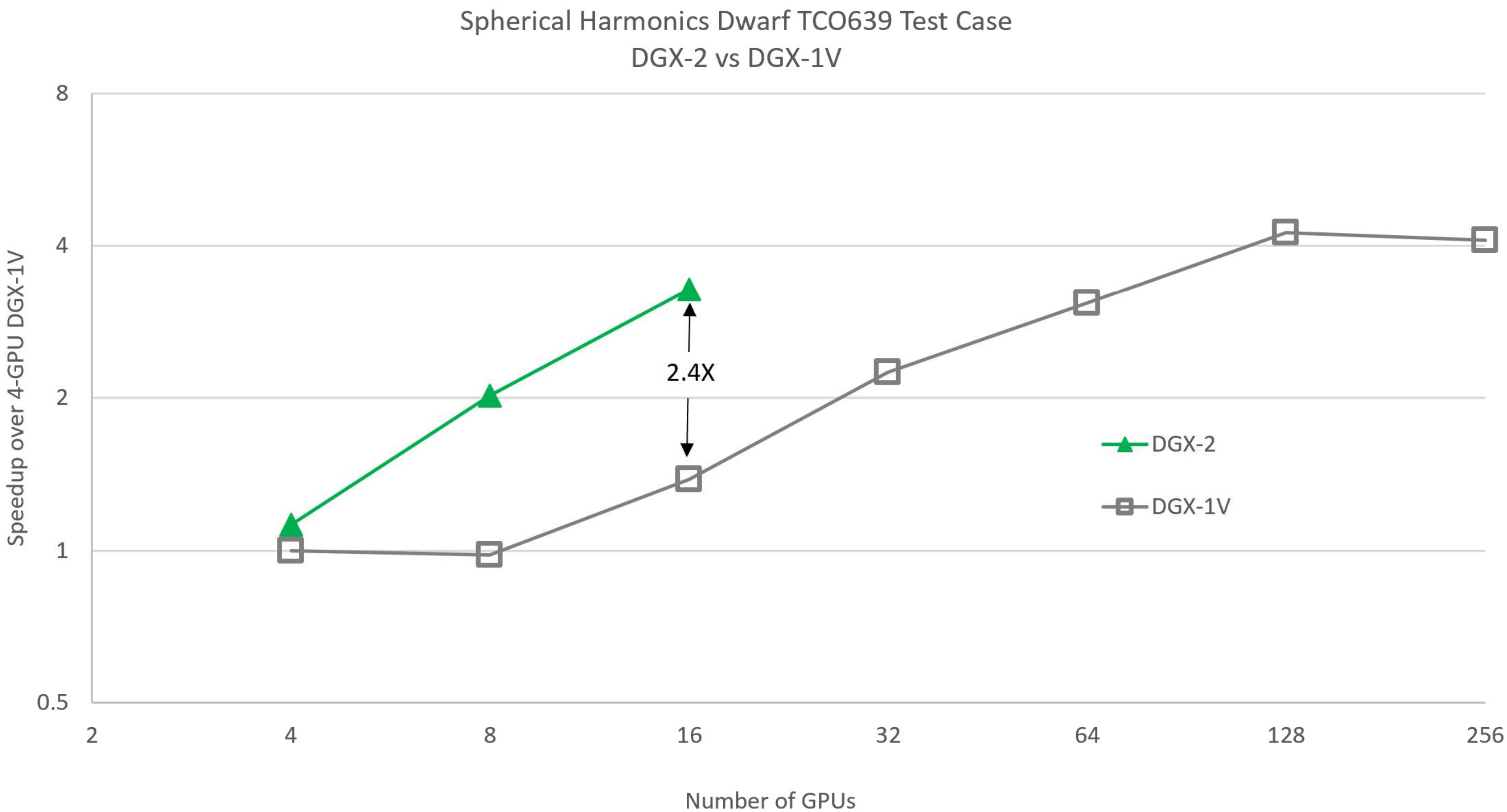
- When using all 8 GPU in DGX-1V:
  - No AlltoAll NVLINK Connectivity – some messages go through PCIe and system memory
  - This limits performance
- When using 16 GPUs across 2 DGX-1V servers
  - Some messages go across Infiniband network
  - Further bottleneck

DGX-2 WITH NVSWITCH



- AlltoAll network architecture with NVSwitch maps perfectly to the problem.
- Full bandwidth between each GPU pair.

SPHERICAL HARMONICS: DGX-2 VS DGX-1V



DGX-1V uses MPI for >=8 GPUs (due to lack of AlltoAll links), all others use CUDA IPC.
DGX-2 results use pre-production hardware.

SUMMARY

- Optimizing the exposure of parallelism, memory coalescing and data management can have dramatic effects on performance.
- SH single-GPU performance is vastly improved, but FFT part remains sub-optimal.
  - Implementation of batching where different sizes are allowed within each batch would expectedly fix this.
- DGX-2/NVSwitch all-to-all connectivity allows SH to scale to all 16 GPUs.
- MPDATA single-GPU performance is now optimal.
- MPDATA multi-GPU has also been optimized. Data volume involved in exchange is less than for SH, so scaling is better on DGX-1V (but not ideal). Still to perform MPDATA experiments on DGX-2.
- These results give indications that multi-GPU systems can be effectively exploited to allow forecasting agencies to continue to further improve weather predictions.

NVIDIA ACTIVE COLLABORATIONS ON ATMOSPHERE MODELS

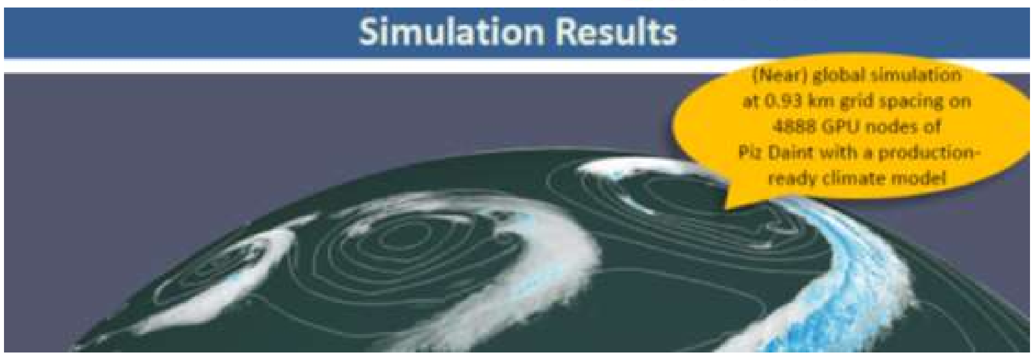
	Model	Organisations	Funding Programme
Global	E3SM-Atm, SAM	DOE: ORNL, SNL	E3SM, DOE ECP
	MPAS-A	NCAR, UWyo, KISTI, IBM	WACA II
	FV3/UFS	NOAA	NOAA SENA
	NUMA/NEPTUNE	US Naval Res Lab, NPS	ONR / NPS
	IFS	ECMWF	ESCAPE
	GungHo/LFRic	MetOffice, STFC	PSyclone
	ICON	DWD, MPI-M, CSCS, MCH	PASC ENIAC
	KIM	KIAPS	KMA
Regional	COSMO	MCH, CSCS, DWD	PASC GridTools
	WRFg	NVIDIA, NCAR	*None / NVIDIA*
	AceCAST-WRF	TempoQuest	Venture backed

LARGE SCALE ATMOSPHERE AT ~1KM: COSMO

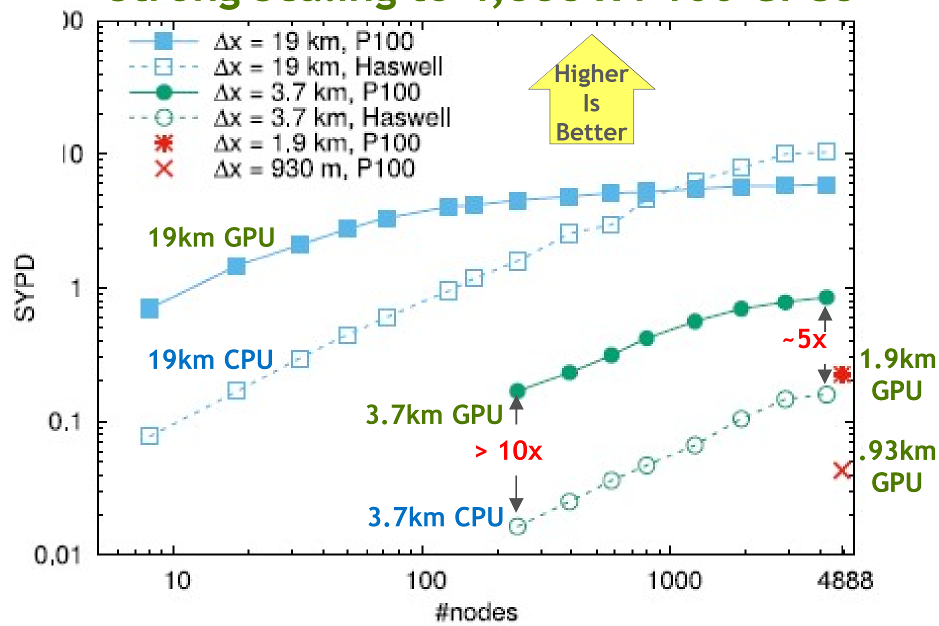
Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4'888 GPUs with COSMO 5.0

Oliver Fuhrer[1], Tarun Chadha[2], Torsten Hoefler[3], Grzegorz Kwasniewski[3], Xavier Lapillonne[1], David Leutwyler[4], Daniel Lüthi[4], Carlos Osuna[1], Christoph Schär[4], Thomas C. Schulthess[5,6], and Hannes Vogt[6]

- Oliver Fuhrer, et al, MeteoSwiss



Strong Scaling to 4,888 x P100 GPUs



Piz Daint
#6 Top500
25.3 PetaFLOPS
5320 x P100 GPUs

Δx	# nodes	SYPD	MWh / SY
0.93 km	4,888	0.043	596
1.9 km	4,888	0.23	97.8

Source: https://www.geosci-model-dev-discuss.net/gmd-2017-230/

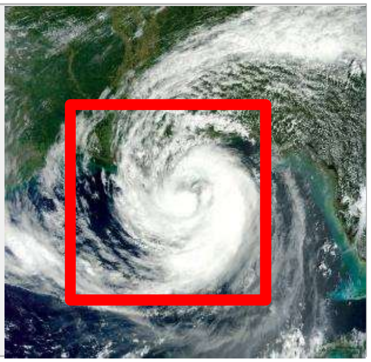
DEEP LEARNING APPLICATIONS IN CLIMATE AND WEATHER

-View Poster in Weather Room by Dr. D. Hall, NVIDIA

DETECTION

- Tropical storms

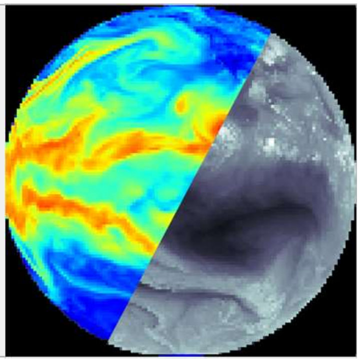
Extra-tropical cyclones
Atmospheric rivers
Cyclogenesis events
Convection initiation
Change detection



TRANSLATION

- Data Assimilation
- Satellite Emulation

Model inter-comparison
Common data formatting
Colorization

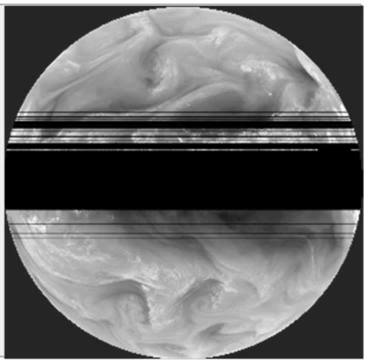


ENHANCEMENT

Frame repair
Sequence repair

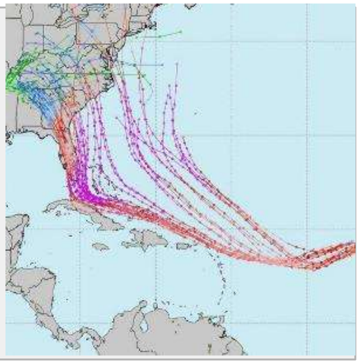
- Slow motion

Anomaly detection
Super-resolution
Cloud removal



PREDICTION

Uncertainty prediction
Storm track
Storm intensity
Fluid motion
Now casting



EMULATION

Physical parametrizations
Turbulence
Radiation
Convection
Solver Acceleration