GPU Developments for the NEMO Model

Stan Posey, HPC Program Manager, ESM Domain, NVIDIA (HQ), Santa Clara, CA, USA

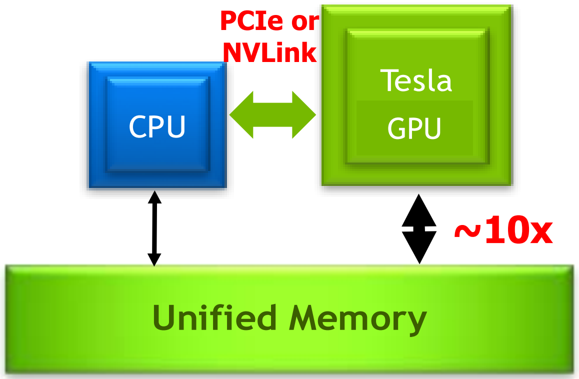
NVIDIA.

TOPICS OF DISCUSSION

- NVIDIA HPC AND ESM UPDATE
- GPU PROGRESS ON NEMO MODEL

NVIDIA GPU Introduction

GPU Introduction



- Co-processor to the CPU
- Threaded Parallel (SIMT)
- CPUs: x86 | Power
- HPC Motivation:
  - Performance
  - Efficiency
  - Cost Savings

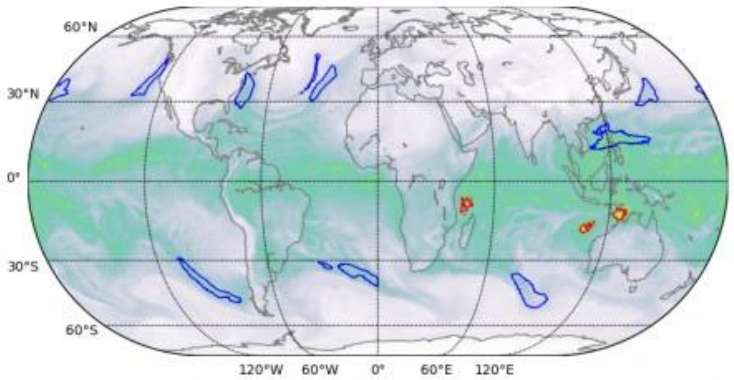
TOP 500 The List. **#1 Summit:**

US DOE Oak Ridge NL

4,600 nodes

IBM Power9 CPUs

27,600 x NVIDIA GPUs



2018 Gordon Bell Finalist:

"AI and the Summit GPU Accelerated Supercomputer Helps Identify Extreme Weather Patterns"

http://cs.lbl.gov/news-media/news/2018/berkeley-lab-oak-ridge-nvidia-team-breaks-exaop-barrier-with-deep-learning-application/

NVIDIA GPUs in the Top 10 of Current Top500

"It's somewhat ironic that training for deep learning probably has more similarity to the Top500 HPL benchmark than many of the simulations that are run today…"
- - Kathy Yelick, LBNL, USA

TOP 500 The List. June 2018

- Top500 #1: ORNL Summit
  - RMAX (HPL) = 122 PF
  - IBM Power + GPU system
- Top #1's in 3 regions with GPUs:
  - USA - ORNL Summit
  - Europe - CSCS Piz Daint
  - Japan - AIST ABCI
- Total 5 of Top7 with GPUs
- #1 Summit vs. #7 Titan (2012)
  - ~7x more performance
  - ~Same power consumption

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	187,659.3	8,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
3	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/NNSA/LLNL United States	1,572,480	71,610.0	119,193.6	
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649
6	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
7	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209

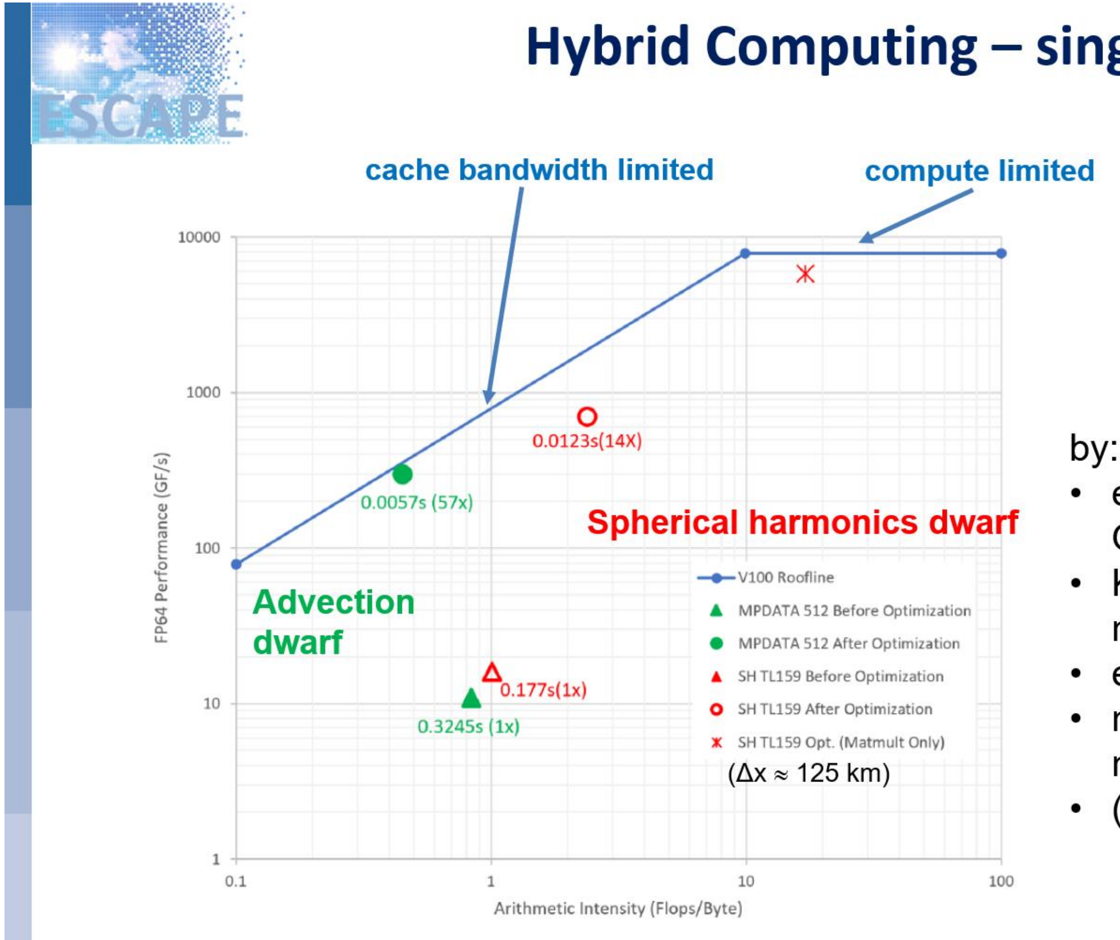
NVIDIA

NVIDIA GPU Collaborations on Atmospheric Models

Model	Organizations	Funding Source
E3SM-Atm, SAM	DOE: ORNL, SNL	E3SM, ECP
MPAS-A	NCAR, UWyo, KISTI, IBM	WACA II
FV3/UFS	NOAA	SENA
NUMA/NEPTUNE	NPS, US Naval Res Lab	NPS
IFS	ECMWF	ESCAPE
GungHo/LFRic	MetOffice, STFC	PSyclone
ICON	DWD, MPI-M, CSCS, MCH	PASC ENIAC
KIM	KIAPS	KMA
COSMO	MCH, CSCS, DWD	PASC GridTools
WRFg	NVIDIA, NCAR	*NVIDIA*
AceCAST-WRF	TempoQuest	Venture backed

ECMWF IFS Spherical Harmonic Dwarf - Single-GPU

Hybrid Computing – single GPU



SH Dwarf = 14x

Advection = 57x

by:
- exposing parallelism in loops for OpenACC mapping
- Kernel optimization by memory mapping
- exploiting CUDA BLAS features
- minimizing data allocation and movement
- (calling C/CUDA from PGI Fortran)

From *"ECMWF Scalability Programme"*

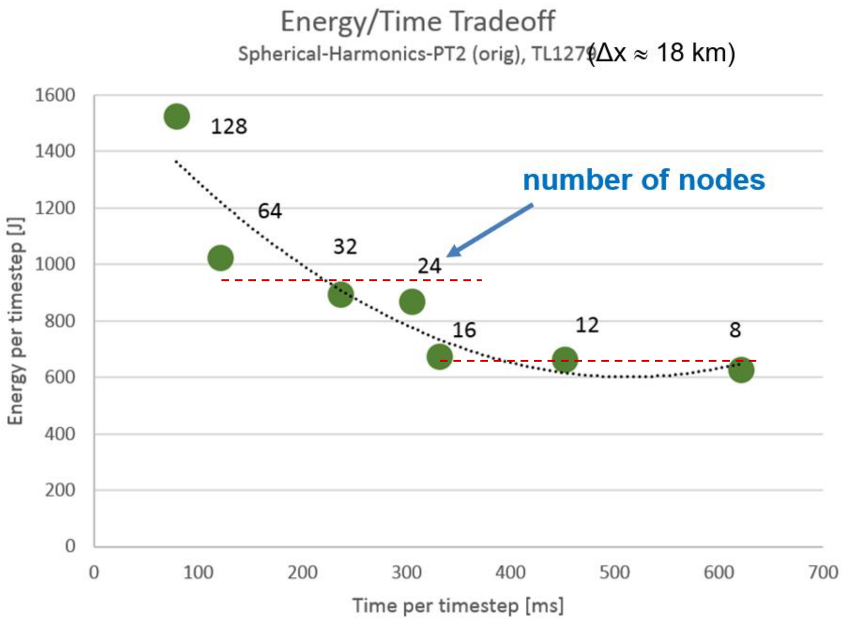
Dr. Peter Bauer,
UM User Workshop,
MetOffice, Exeter, UK
15 June 2018

- Single V100 GPU improved SH dwarf by 14x
- Single V100 GPU improved MPDATA dwarf by 57x

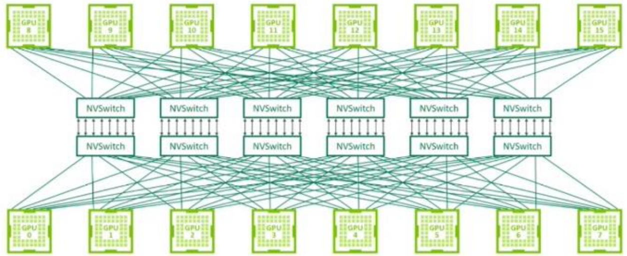
ECMWF MeteoSwiss dmi RMI NVIDIA METEO FRANCE Toujours un temps d'avance ICHEC Deutscher Wetterdienst Optalysys Bull PSNC Loughborough University

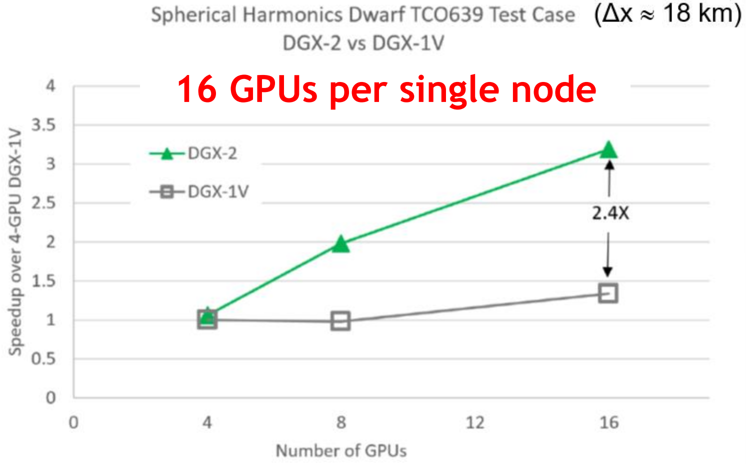
ECMWF IFS Spherical Harmonic Dwarf - Multi-GPU

Hybrid Computing – multiple GPU



NVIDIA NGX-2 with NVSwitch





From *"ECMWF Scalability Programme"*

Dr. Peter Bauer,
UM User Workshop,
MetOffice, Exeter, UK
15 June 2018

- Results of Spherical Harmonics Dwarf on NVIDIA DGX System
- Additional 2.4x gain from DGX-2 NVSwitch for 16 GPU systems

ECMWF MeteoSwiss RMI NVIDIA METEO FRANCE ICHEC Deutscher Wetterdienst Optalysys Bull PSNC Loughborough University

TOPICS OF DISCUSSION

- NVIDIA HPC AND ESM UPDATE
- **GPU PROGRESS ON NEMO MODEL**

Programming Strategies for GPU Acceleration

Applications		
GPU Libraries	**OpenACC Directives**	**Programming in CUDA**
Provides Fast "Drop-In" Acceleration	GPU-acceleration in Standard Language (Fortran, C, C++)	Maximum Flexibility with GPU Architecture and Software Features

Increasing Development Effort →

NOTE: Many application developments include a combination of these strategies

Examples

- IFS: FFT, DGEMM
- COSMO: Tridiag Solve

- FV3
- MPAS
- IFS
- ICON
- COSMO (Physics)
- WRF
- E3SM (ACME)
- CAM-SE
- NICAM
- NEMO
- LFRic/Gungho

- COSMO (Dycore)
- NUMA

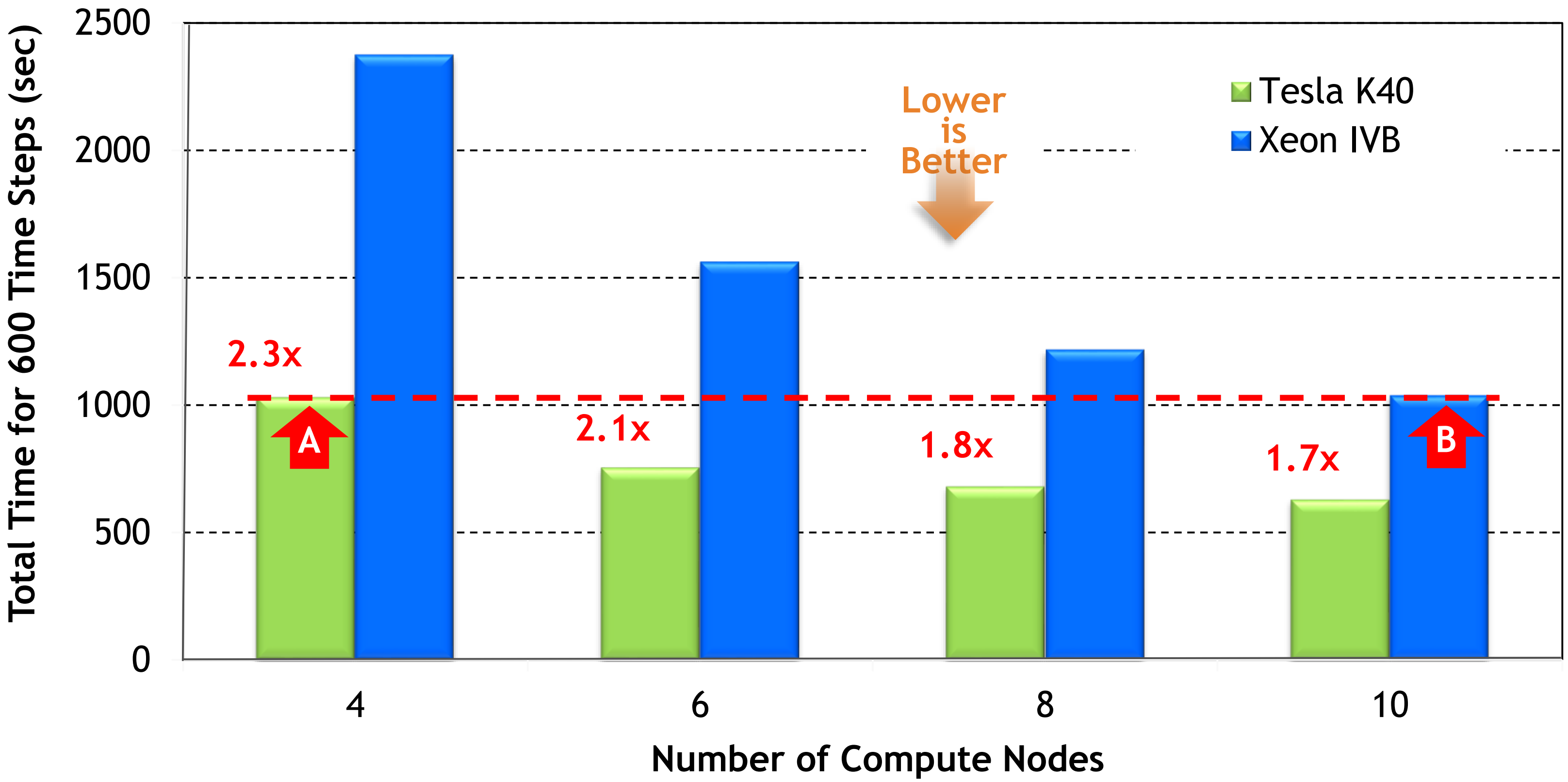
Early GPU Developments of NEMO (2014)

- **NVIDIA-led investigations during 2014 by Dr. Jeremy Appleyard, NVIDIA UK:**
  - www.fz-juelich.de/SharedDocs/Downloads/IAS/JSC/EN/slides/nvidia-ws-2014/04-appleyard-nemo.pdf?__blob=publicationFile

- **OpenACC conclusion the most natural GPU solution for NEMO**
  - **CPU profiles for v3.5 code were flat, no hotspots available for quick acceleration**
  - **Portable solution: OpenACC available on NVIDIA and AMD GPUs, x86 and Power CPUs**
  - **OpenACC directives offer ease of maintenance with existing NEMO Fortran code**

- **Investigations based on NEMO v3.5 rev 3903 development code (unreleased)**
  - **GPU implementation using OpenACC to minimize code changes among other benefits**
  - **Final OpenACC + Fortran code was (nearly) only insertion of OpenACC directives**
    - **Changes to MPI routines to 'batch' multiple sequential calls**
    - **MPI modifications were also beneficial to CPU-only code**
  - **OpenACC code ran correctly on CPUs when not invoking OpenACC flag at compile time**
  - **OpenACC approach, performance results, and conclusions summarized on next slides**

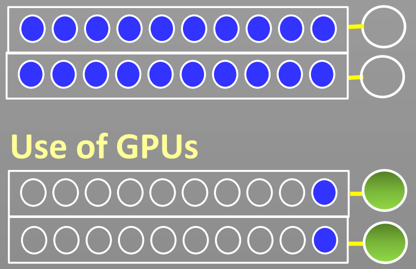
Strong Scaling for ORCA025 Configuration (2014)



Single node utilization:
2 x IVB + 2 x K40

Without using GPUs

Use of GPUs

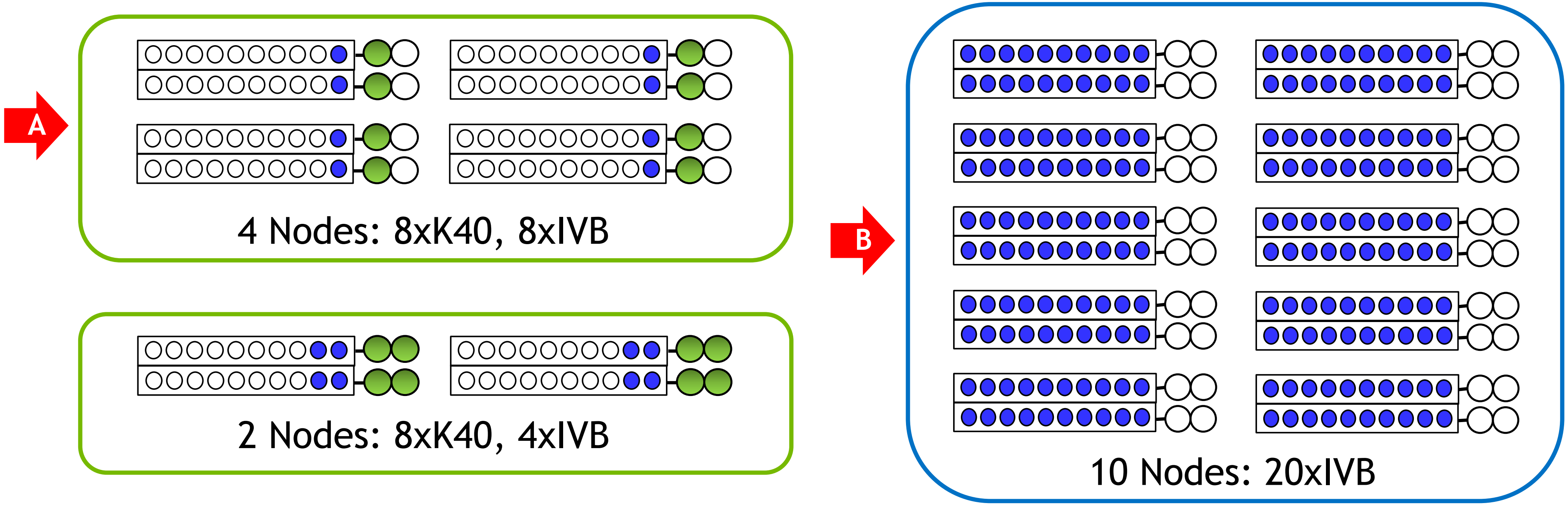


ORCA025 settings:

- Output every 5 days
- Total run: 10 days
- Time steps: 600
- LIM2 ice model
- Uniform horizontal grid of 1442 x 1021
- Variable vertical grid of 75 levels

Strong Scaling of ORCA025 Configuration (2014)

NODE Configurations with Same Performance for ORCA025



HPC TRENDS: Node configurations in 2018 are 6+ GPUs:

- MeteoSwiss operational NWP system = 8 x K80 per node
- NOAA weather/climate research system = 8 x P100 per node
- ORNL Summit system for E3SM model = 6 x V100 per node
- ECMWF IFS spherical harmonic dwarf study = 16 x V100 per node

Feature Progression in NVIDIA GPU Architectures

	V100 (2017)	P100 (2016)	K40 (2014)
Double Precision TFlop/s	7.5 1.4x - 5.4x	5.3 3.8x	1.4
Single Precision TFlop/s	15.0 1.4x - 3.5x	10.6 2.5x	4.3
Half Precision TFlop/s	120 (DL) ~6x	21.2	n/a
Memory Bandwidth (GB/s)	900 1.25x - 3.1x	720 2.5x	288
Memory Size	16 or 32GB 2.00x	16GB 1.33x	12GB
Interconnect	NVLink: Up to 300 GB/s PCIe: 32 GB/s	NVLink: 160 GB/s PCIe: 32 GB/s	PCIe: 16 GB/s
Power	250W - 300W	300W	235W

V100 Availability

CRAY DELL EMC Hewlett Packard Enterprise IBM SUPERMICRO

Current GPU Developments of NEMO (2018)

- **Collaboration on 2 thrusts between Met Office, STFC, and NVIDIA/PGI**
  - **#1:** Hand development and optimization of OpenACC code led by MetO with NVIDIA/PGI
    - **Update:** Clean compile of 3.5 OpenACC code (2014) on new V100 GPU with new PGI version
  - **#2:** PSyclone auto-generated OpenACC code, led by STFC with MetO and NVIDIA/PGI
    - **Update:** STFC OpenACC code generation now working for the 'GOcean' API in PSyclone
  - **Idea:** Hand-optimized OpenACC code to provide 'target' code for Psyclone approach
    - **PSyclone-approach proposed by STFC, Andy Porter from HPC WG minutes**

      http://forge.ipsl.jussieu.fr/nemo/wiki/WorkingGroups/NEMO_HPC/Mins_sub_2017_06_13

      http://forge.ipsl.jussieu.fr/nemo/wiki/WorkingGroups/NEMO_HPC/Mins_2017_07_28

      http://forge.ipsl.jussieu.fr/nemo/wiki/WorkingGroups/NEMO_HPC/Mins_sub_2017_10_16

- **Next step: Implement/Migrate OpenACC directives into NEMO 4.0 source**
  - Development teams have begun inspection of 4.0 code for OpenACC directives

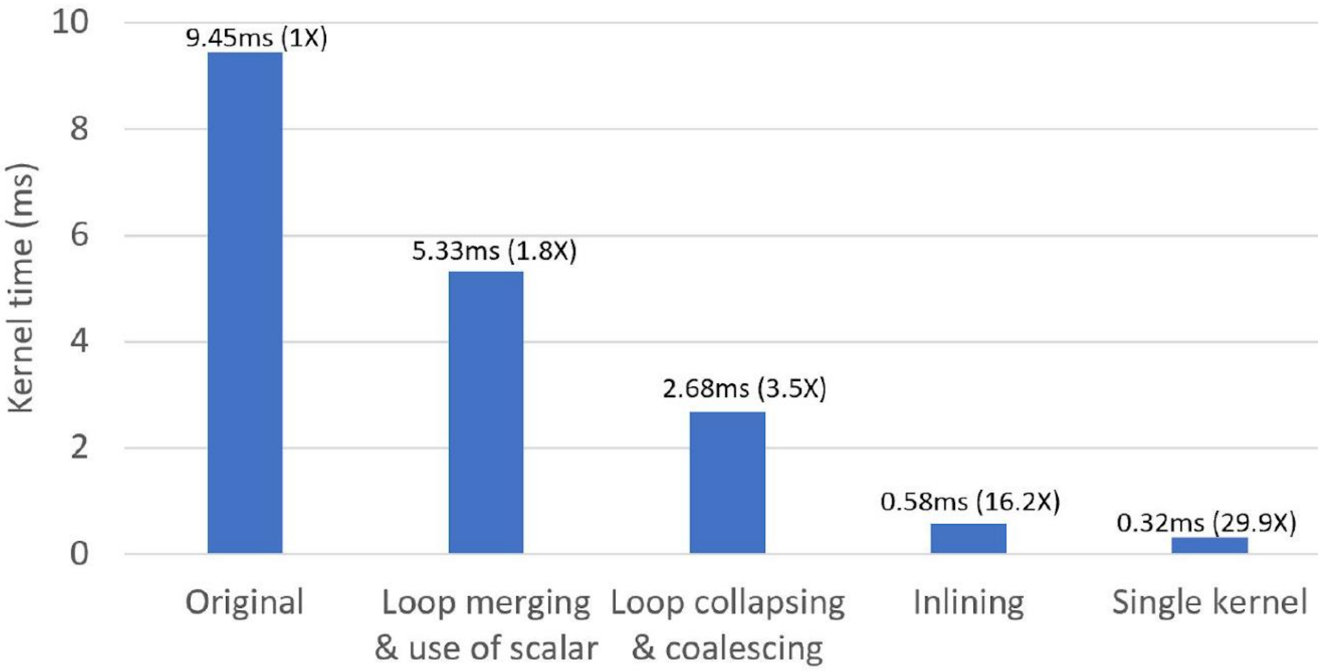
Example: OpenACC Development for LFRIC Model

- **OpenACC collaboration with MetOffice and SFTC: LFRic model**
  - GungHo-MV (matrix-vector operations) OpenACC kernel developed by MetOffice
  - NVIDIA optimizations applied to the OpenACC kernel achieved 30x improvement!
  - Improved OpenACC code provided to STFC as 'target' for Psyclone auto-generation

"Optimization of an OpenACC Weather Simulation Kernel"

- A. Gray, NVIDIA

30x Improvement from NVIDIA Optimizations



NEMO Model and HPC Outlook

- **Good potential exists for a NEMO development on latest GPUs**
  - The V100 GPU has ~3x greater bandwidth vs. the K40 from 2014 results

- **A successful GPU implementation must minimize NEMO code changes**
  - Investigating non-invasive approach PSyclone auto-generation of GPU code

- **Future HPC systems will migrate to heterogeneous architectures**
  - It's time to begin preparation of NEMO code and users to new HPC methods

Thank you and Questions?

Stan Posey, sposey@nvidia.com