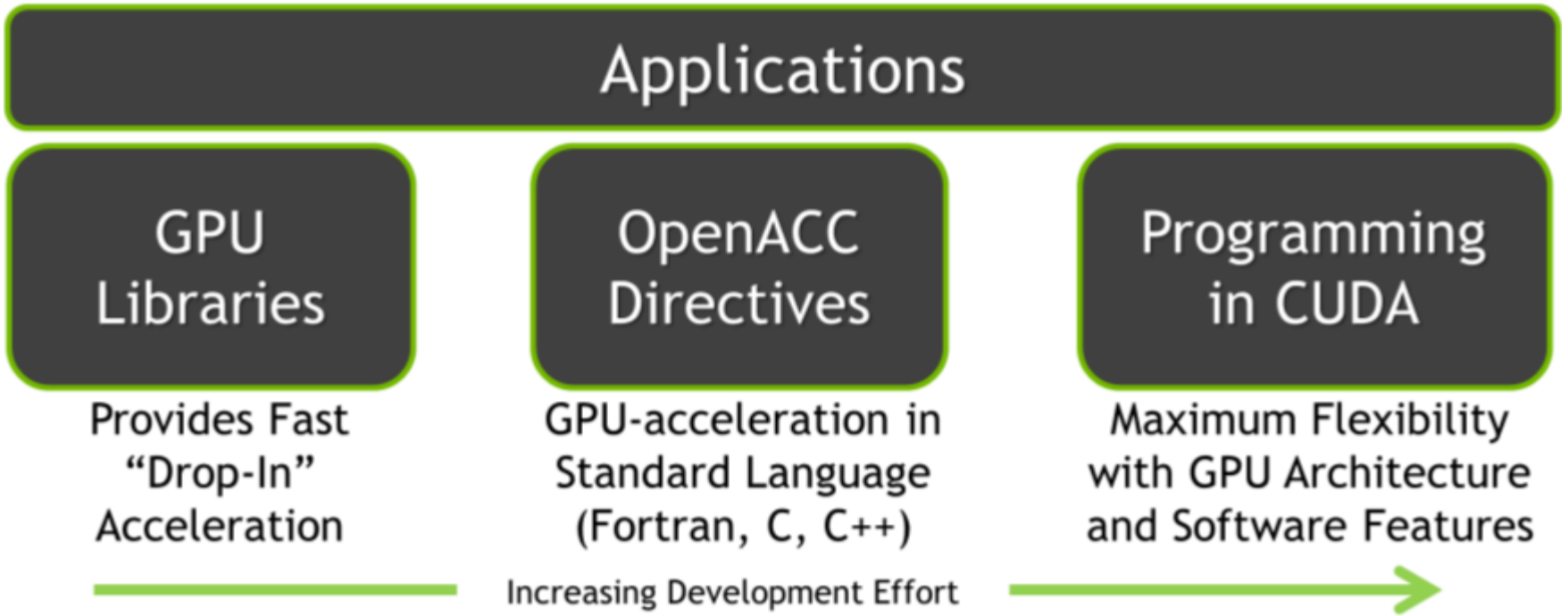
NVIDIA Update for the NEMO HPC WG

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

Summary of NEMO OpenACC Experiments

- **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
- **Project based on NEMO v3.5 rev 3903 development code (unreleased)**
  - **GPU implementation using OpenACC to minimize code changes among other benefits**
  - **OpenACC approach, performance results, and conclusions summarized on next slides**

Programming Strategies for GPU Acceleration



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
- ICON
- NEMO
- UM/Gungho

- COSMO (Dycore)
- NUMA
- ICON
- WRF -TQI
- NICAM (Dycore)

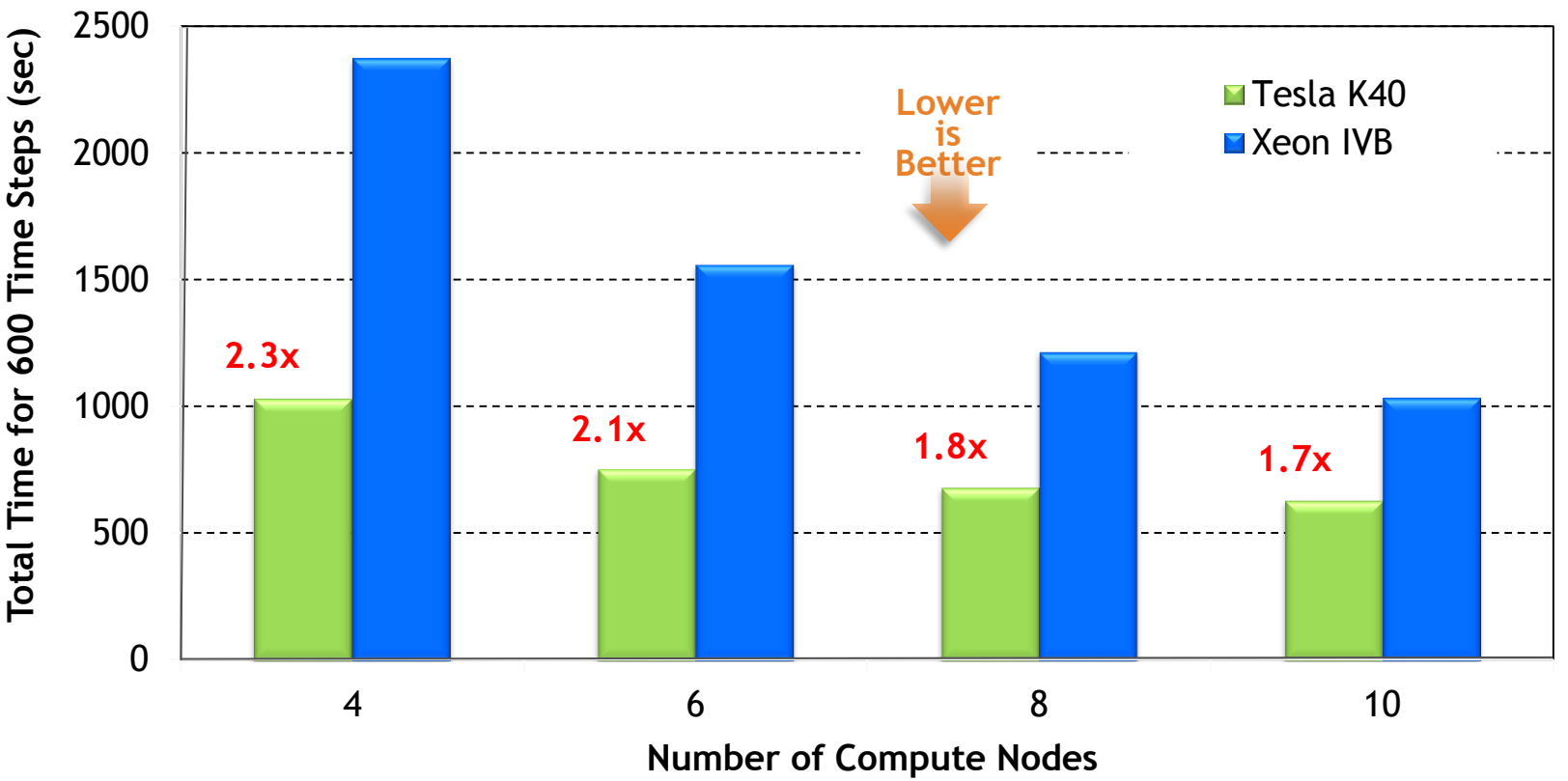
OpenACC Considerations for NEMO

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

- **NVIDIA approach sought to minimize code impact**
  - **Final code was only (nearly) insertion of OpenACC directives**
    - Changes to MPI routines to 'batch' multiple sequential calls
    - MPI modifications were also beneficial to CPU-only code
  - **Code ran correctly on CPUs without invoking OpenACC at compile time**

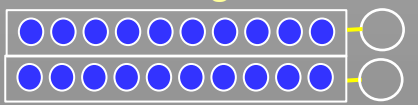
- **OpenACC most common GPU approach for Earth system models**
  - **SC17 Press Release: OpenACC Widely Deployed in Weather and Climate Domain**
    - https://www.openacc.org/news/press-release-openacc-widely-deployed-weather-and-climate-domain
    - Describes achievements of models: **MPAS, IFS (ESCAPE), NIM, COSMO, CAM-S**
  - **PGI Community Edition (freely available) release started with 17.10 (Nov 2017)**
    - http://www.pgroup.com/products/community.htm
    - Free download; Support for V100; Support for unified memory; ++

Strong Scaling for ORCA025 Configuration

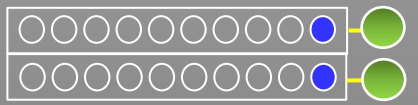


Single node utilization:
2 x IVB + 2 x K40

Without using GPUs



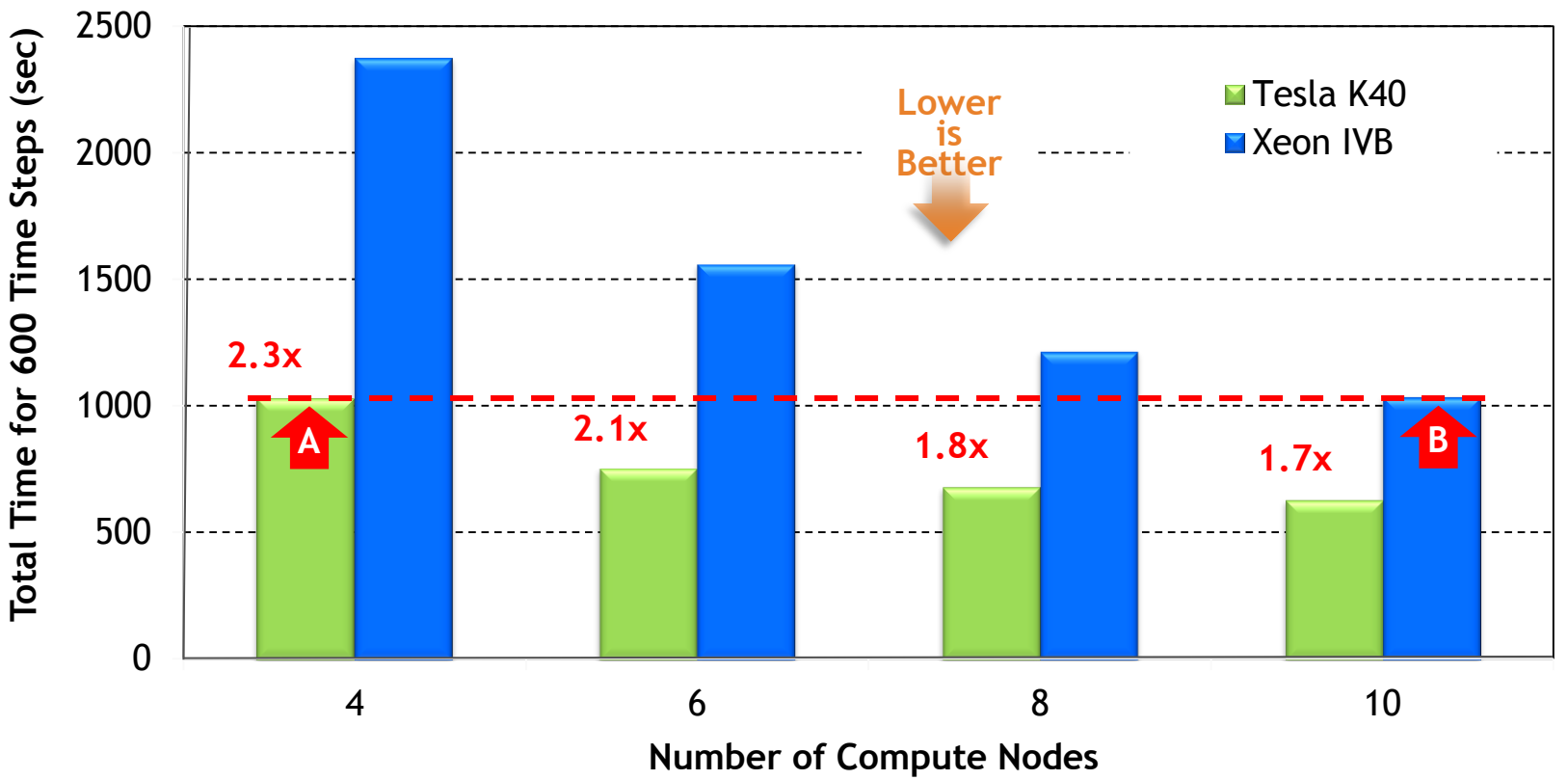
Use of GPUs



ORCA025 settings:

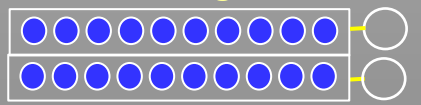
- Output every 5 days
- Total run: 10 days
- Time steps: 600
- LIM2 ice model
- Uniform horizontal grid of 1442 x 1021
- Variable vertical grid of 75 levels

Strong Scaling for ORCA025 Configuration

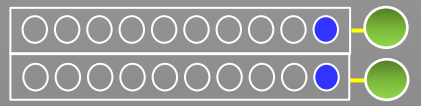


Single node utilization:
2 x IVB + 2 x K40

Without using GPUs



Use of GPUs



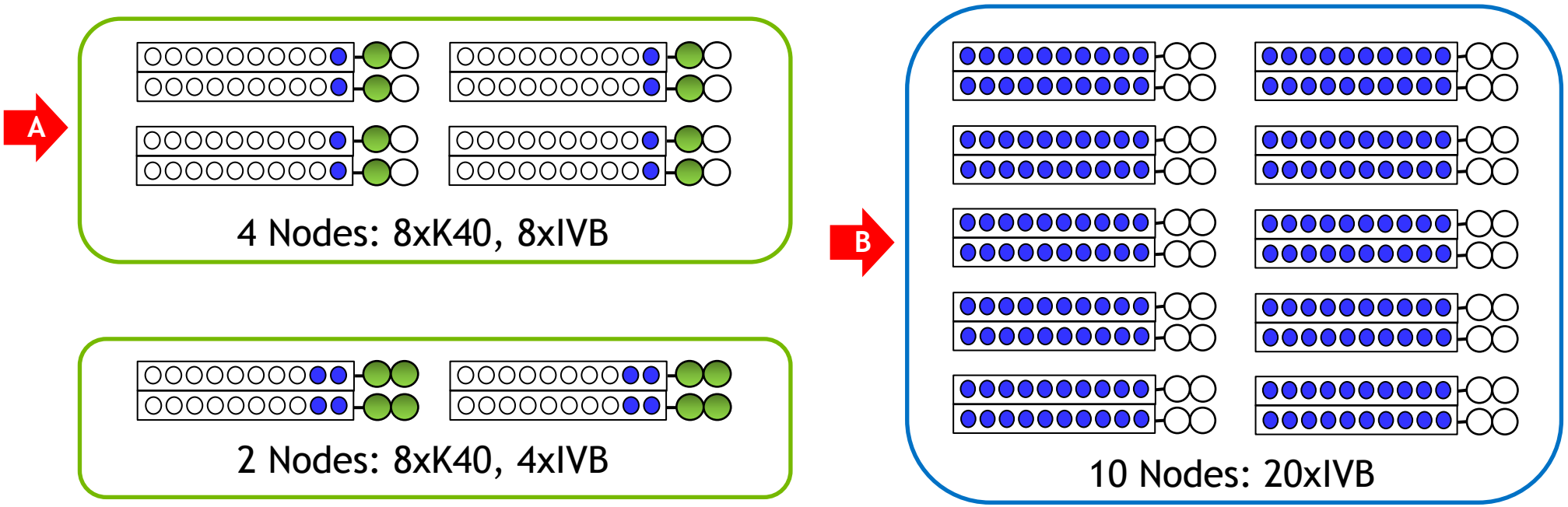
ORCA025 settings:

- 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

NODE Configurations

Configurations that give similar performance for ORCA025



NOTE: Most node configurations today are 4+ GPUs; Examples: MeteoSwiss operational system for COSMO has 8 x K80 per node, NOAA research system has 8 x P100 per node

GPU NEMO Project Conclusions and Outlook

- **Conclusions**
  - **Performance reasonable but not fully explored**
    - Schedule did not permit OpenACC tuning, and/or exploration of compiler optimizations
    - Restrictions on code refactoring limit some identified performance opportunities
  - **Strong scaling limitations in SOR linear solver**
    - Lots of communication from lots of very small kernels leads to inefficient use of the GPU
    - Solution from more efficient MPI packing/unpacking; reduced solver communications
  - **New OpenACC features would benefit**
    - Use of a new feature for unified memory space would simplify porting and debugging
    - This implementation relied on implicit copies and incremental enabling of data regions (pcopy)
- **Outlook**
  - **Significant boost expected from current GPU architecture**
    - New “Volta” V100 offers 3.4x more memory bandwidth vs. K40 (V100 details on next slide)
    - Neglecting communication memory bandwidth was the main limiter in this implementation
    - NVLink is over 3x faster than PCI-E for significantly faster communications within a node

NVIDIA V100 vs. K40 Performance Improvements

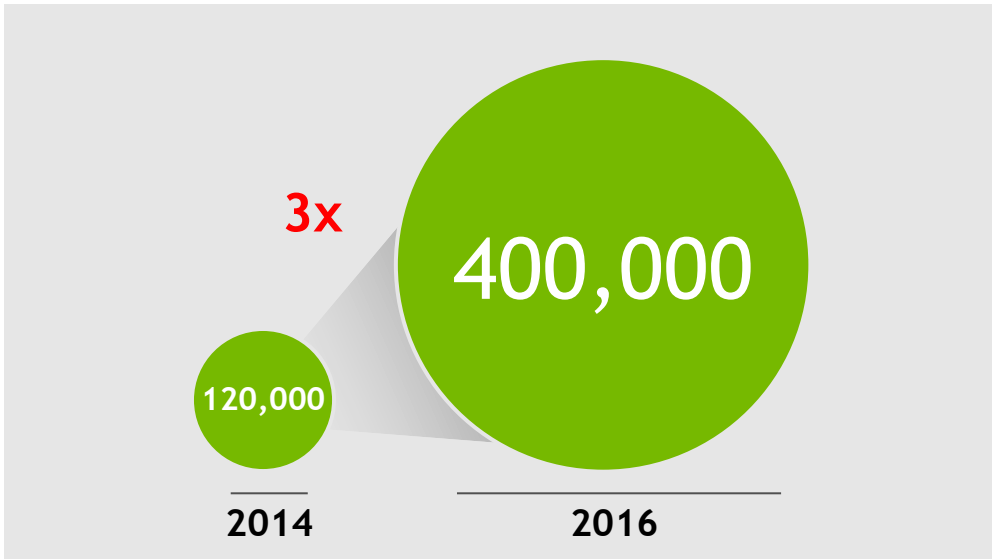
	V100 (2017)		P100 (2016)		K40 (2014)
Double Precision TFlop/s	7.5	1.4x - 5.4x	5.3	3.8x	1.4
Single Precision TFlop/s	15.0	1.4x - 3.5x	10.6	2.5x	4.3
Half Precision TFlop/s	120 (DL)	~6x	21.2		n/a
Memory Bandwidth (GB/s)	990	1.3x - 3.4x	720	2.5x	288
Memory Size	16GB	1.00x	16GB	1.33x	12GB
Interconnect	NVLink: Up to 300 GB/s PCIe: 32 GB/s		NVLink: 160 GB/s PCIe: 32 GB/s		PCIe: 16 GB/s
Power	300W	1.00x	300W		235W

V100 Availability

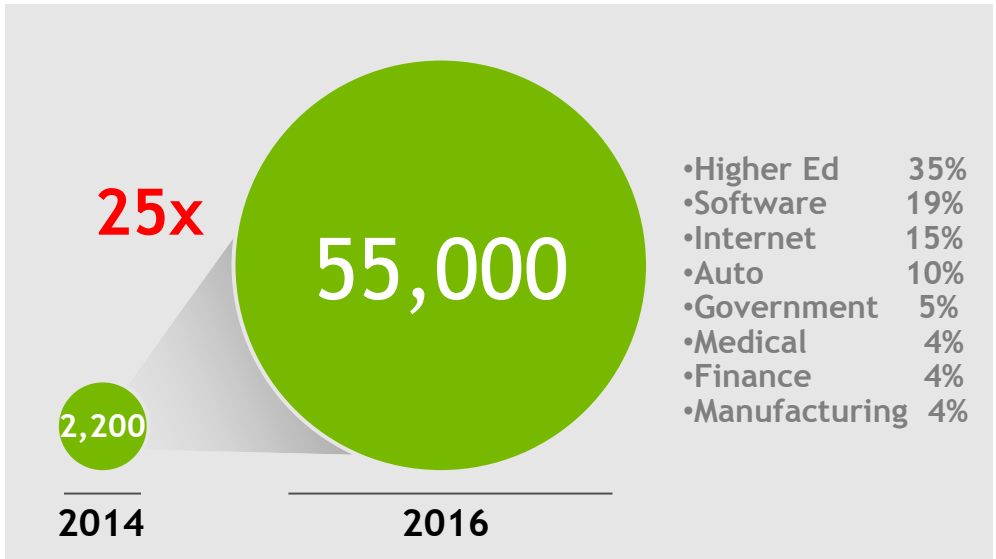
DGX-1: Q3 2017; **OEM : Q4 2017**

DELL EMC, Hewlett Packard Enterprise, IBM, SUPERMICRO

NVIDIA HPC and AI Growth 2014 - 2016



3X HPC Developers



25x Deep Learning Developers

Sample of Large Data Centers with GPUs Installed for HPC and AI:

amazon web services · facebook · Google · IBM · Lawrence Livermore National Laboratory · Microsoft · OAK RIDGE National Laboratory · twitter

Summary of NEMO OpenACC Experiments

- 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
- Project based on NEMO v3.5 rev 3903 development code (unreleased)
  - GPU implementation using OpenACC to minimize code changes among other benefits
  - OpenACC approach, performance results, and conclusions summarized on next slides
- **Detailed results and discussion at NEMO HPC WG May 2016, minutes posted at:**
  - http://forge.ipsl.jussieu.fr/nemo/wiki/WorkingGroups/NEMO_HPC/Mins_2016_05_20
  - "Jeremy Appleyard presented. It was noted that the SOR solver is not included in the vn3.6 and future versions of NEMO. The issue of whether OpenACC directives should be moved up to vn3.6 or included in the trunk was raised and briefly discussed. OpenMP directives are currently being ported into the trunk."
- **NVIDIA proposed next steps**
  - **Test existing v3.5 rev 3903 OpenACC code on V100, confirm ~3x improvement vs. K40**
  - **Assess the technical effort to migrate existing OpenACC directives to v3.6 stable release**
  - **Explore collaborations with the NEMO community for next meaningful GPU milestone?**
    - Example: Psyclone-approach proposed by Andy Porter at STFC
      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

Thank you and Questions?

Stan Posey, sposey@nvidia.com