UNIVERSITY OF CAMBRIDGE

Early Experiments using GPU Direct on Wilkes-2, the new Cambridge Petascale GPU-Accelerated Cluster

Filippo SPIGA[1] <fs395@cam.ac.uk>

[1] Head RSE, Research Computing Services, University of Cambridge

History of GPU systems in Cambridge

- 2009: 128 NVIDIA S1070, 120 TFlops/s in single precision
- November 2013: **Wilkes-1**, 256 NVIDIA K20c
  - Designed as “MPP-like machine”: 2 GPU, 2 sockets, dual-rail Mellanox Connect-IB
  - Community: parallel codes, mainly CFD and astro-physics
  - Opportunities: exploring scaling GPU codes (→ MVAPICH2-GDR)
  - Challenges: scheduling, effective node sharing (MPI placement, thread affinity, NUMA effects)
- June 2017: **Wilkes-2**, 360 NVIDIA P100
  - Designed as cluster of “(semi) fat node”: 4 GPU, 1 socket, single-rail Mellanox EDR
  - Communities: parallel codes (CFD), accelerated serial codes, Deep Learning
  - Opportunities: exploring scaling, transparent intra-node GPU-GPU via MPI+GPU
  - Challenges: optimize intra-/inter-node traffic (→ MVAPICH2-GDR)

New UK record

Top500 June 2017

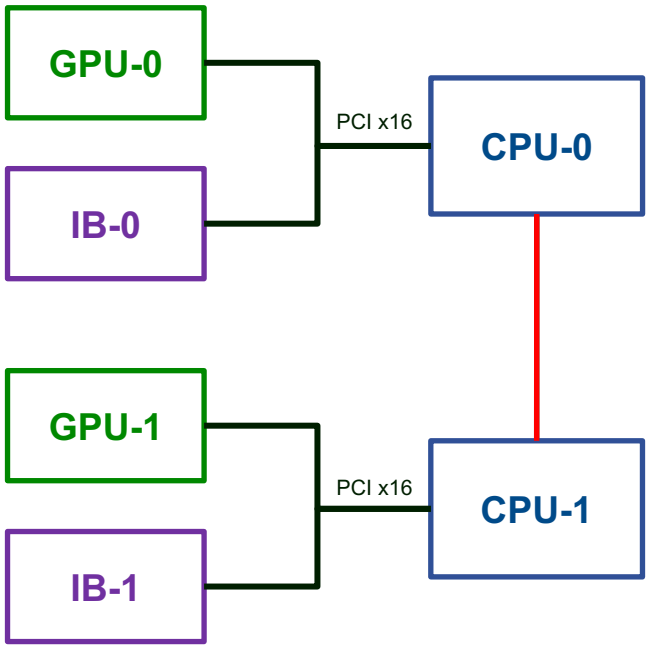
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
97	**SNL/NNSA CTS-1 Cayenne** - Tundra Extreme Scale, Xeon E5-2695v4 18C 2.1GHz, Intel Omni-Path , Penguin Computing Sandia National Laboratories United States	40,392	1,223.7	1,357.2	480
98	**SNL/NNSA CTS-1 Serrano** - Tundra Extreme Scale, Xeon E5-2695v4 18C 2.1GHz, Intel Omni-Path , Penguin Computing Sandia National Laboratories United States	40,392	1,223.7	1,357.2	480
99	Cray XC40, E5-2680v3 12C 2.5Ghz/Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , Cray Inc. Deutscher Wetterdienst Germany	41,472	1,214.2	1,459.8	528.2
100	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Dell University of Cambridge United Kingdom	21,240	1,193.0	1,751.6	114.4
101	**LLNL/NNSA CTS-1 Nel** - Tundra Extreme Scale, Xeon E5-2695v4 18C 2.1GHz, Intel Omni-Path , Penguin Computing Lawrence Livermore National Laboratory United States	39,744	1,179.6	1,335.4	485
102	**Sekirei** - SGI ICE XA, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR , HPE University of Tokyo/Institute for Solid State Physics Japan	38,016	1,178.3	1,520.6	580.9
103	**Garnet** - Cray XE6, Opteron 16C 2.500GHz, Cray Gemini interconnect , Cray Inc. ERDC DSRC United States	150,528	1,167.0	1,505.3	5,569

Green500 June 2017

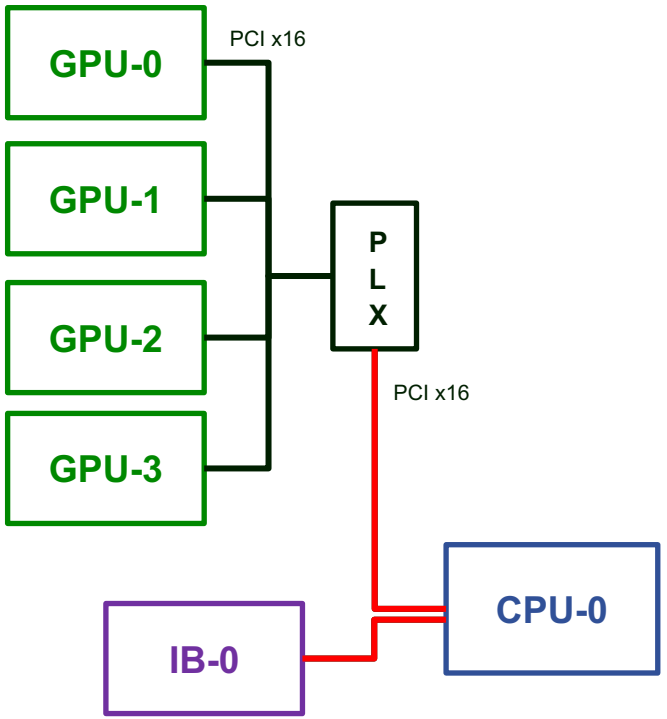
Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	61	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	36,288	1,998.0	142	14.110
2	465	**kukai** - ZettaScaler-1.6 GPGPU system, Xeon E5-2650Lv4 14C 1.7GHz, Infiniband FDR, NVIDIA Tesla P100 , ExaScalar Yahoo Japan Corporation Japan	10,080	460.7	33	14.046
3	148	**AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , NEC National Institute of Advanced Industrial Science and Technology Japan	23,400	961.0	76	12.681
4	305	**RAIDEN GPU subsystem** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Fujitsu Center for Advanced Intelligence Project, RIKEN Japan	11,712	635.1	60	10.603
5	100	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Dell University of Cambridge United Kingdom	21,240	1,193.0	114	10.428
6	3	**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	2,272	10.398
7	69	**Gyoukou** - ZettaScaler-2.0 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , ExaScalar Japan Agency for Marine -Earth Science and Technology Japan	3,176,000	1,677.1	164	10.226

High-level architecture building blocks

Wilkes-1

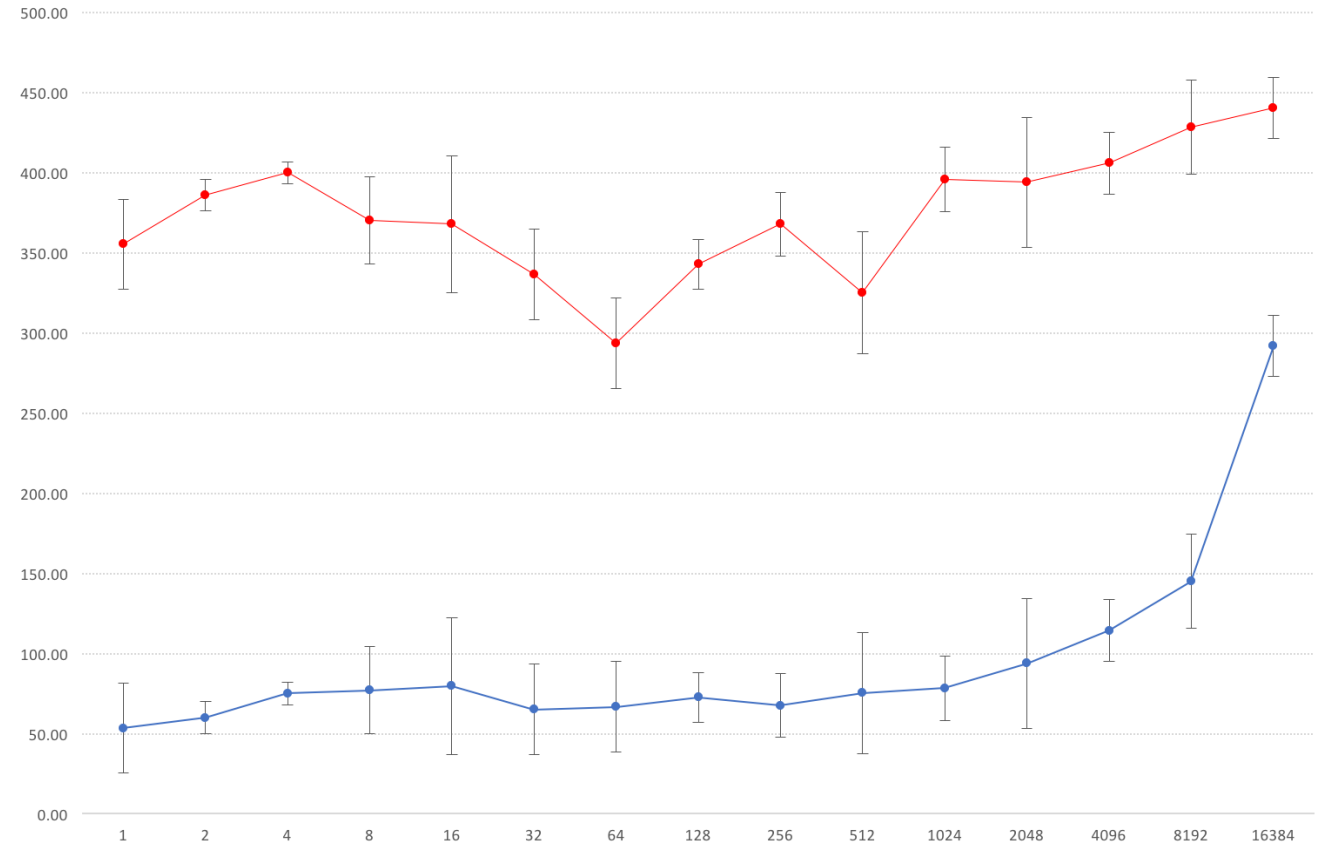


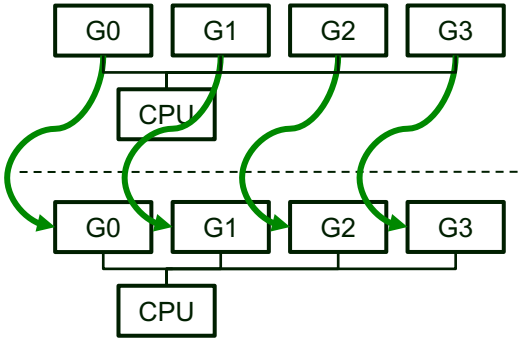
Wilkes-2



Synthetic benchmarks

Q: How vary latency when pairs of MPI communicate together across nodes?





MV2-GDR 2.2-4 “D-D” without GPU Direct

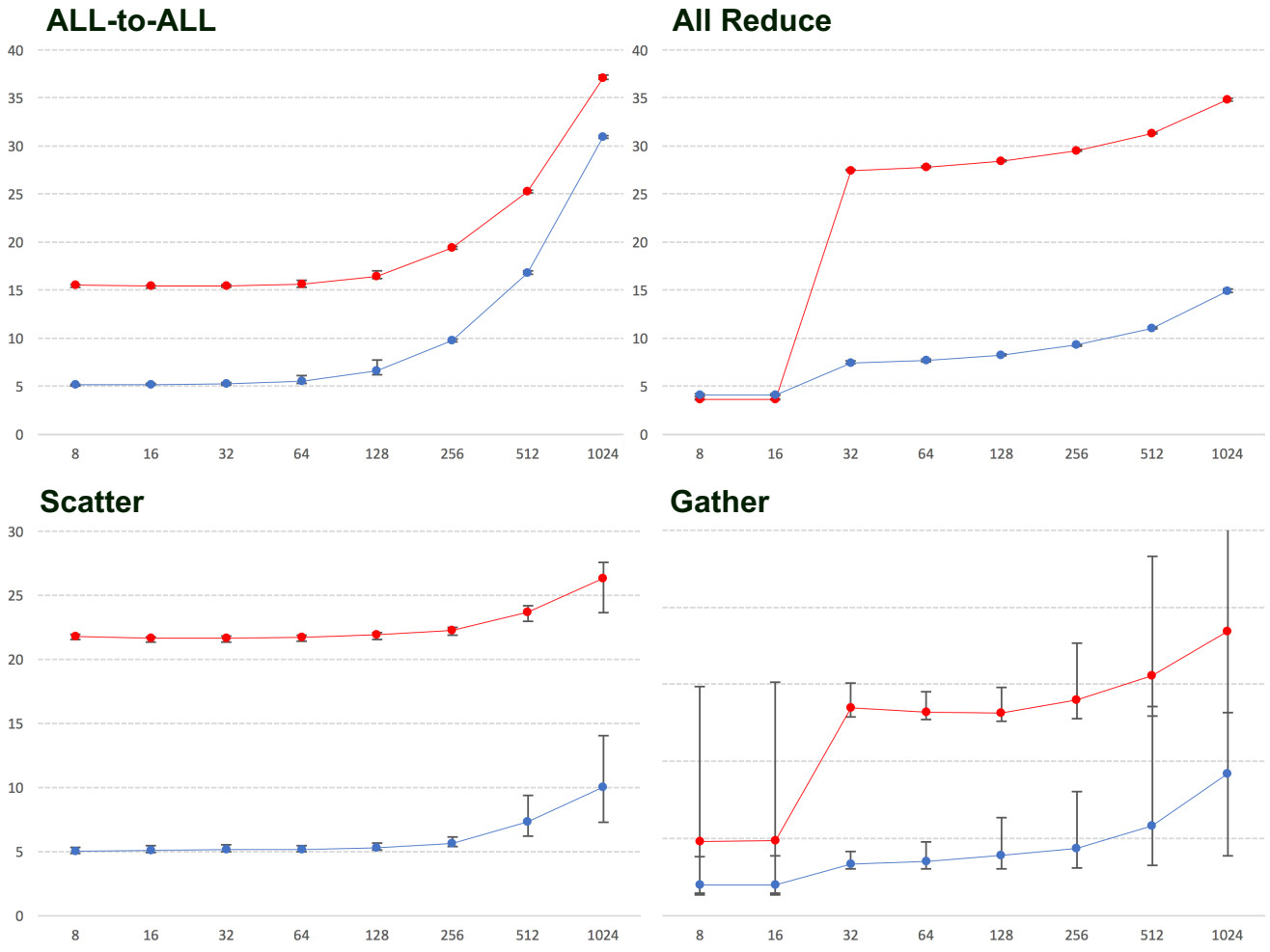
MV2-GDR 2.2-4 “D-D” with GPU Direct

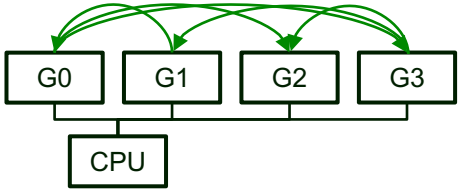
-d cuda -x 1000 -i 2500 D D

Early results, no tuning applied!

Synthetic benchmarks

Q: How vary latency for MPI collectives within the nodes?





MV2-GDR 2.2-4 “D-D” without GPU Direct

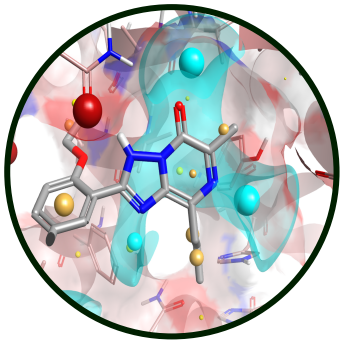
MV2-GDR 2.2-4 “D-D” with GPU Direct

-d cuda -x 1000 -i 2500 D D

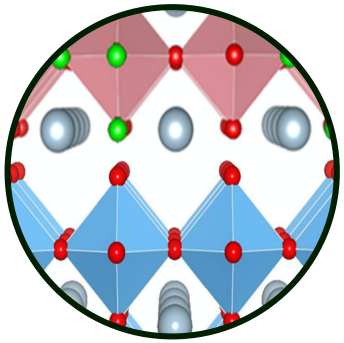
Early results, no tuning applied!

Cambridge Service for Data Driven Discovery (CSD3)

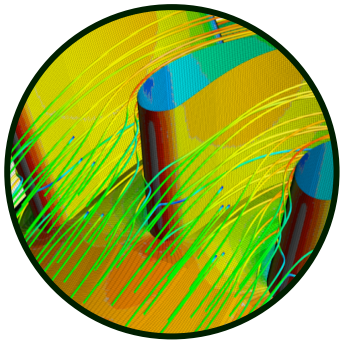
Science Themes



Computational Chemistry



Material Science



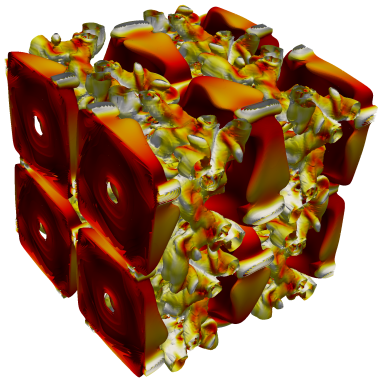
CFD and Combustion

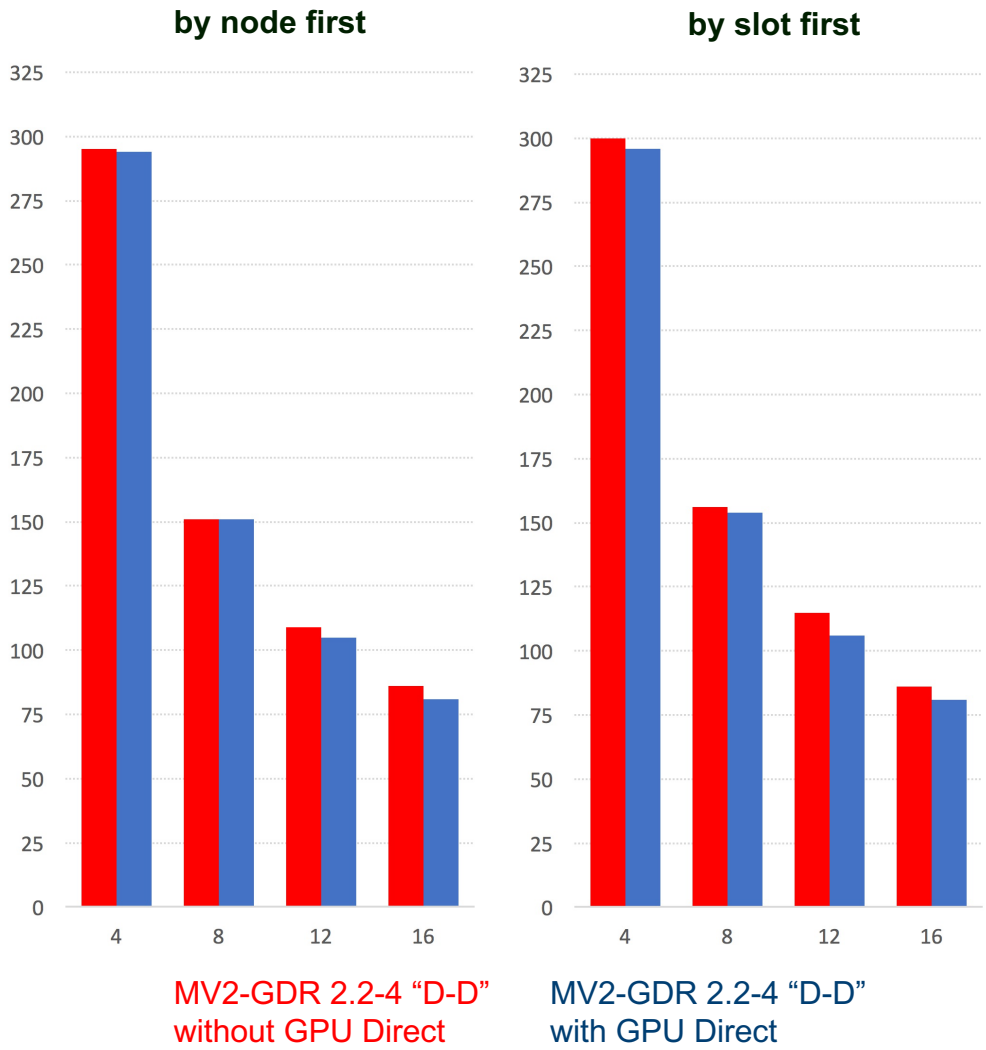
Smart Cities

Health Informatics

PyFR

- Develop at Vincent Lab (ICL)
- CFD via Flux Reconstruction approach of Huynh
- Python + {GPU, CPU}
- Auto-generated code
- Heavy overlap comp/comm
- Gordon Bell finalist 2016





Early results, no tuning applied!

What next?

- Getting ready for production Tier-2 National service in November
- Validate various compilers (including PGI 17.x, very important !)
- Performance tuning / sensible defaults
- Template scripts / documentation (*What* / *Why* / *How*)
- Dissemination via benchmarks

http://csd3.cam.ac.uk

THANK YOU