A CUDA IMPLEMENTATION OF THE HIGH PERFORMANCE CONJUGATE GRADIENT (HPCG) BENCHMARK

Everett Phillips, Massimiliano Fatica

OUTLINE

- CUDA implementation(s) overview
- Single node performance
- Multi node performance
- Comparison to other architectures
- Conclusions/suggestions

CUDA IMPLEMENTATIONS

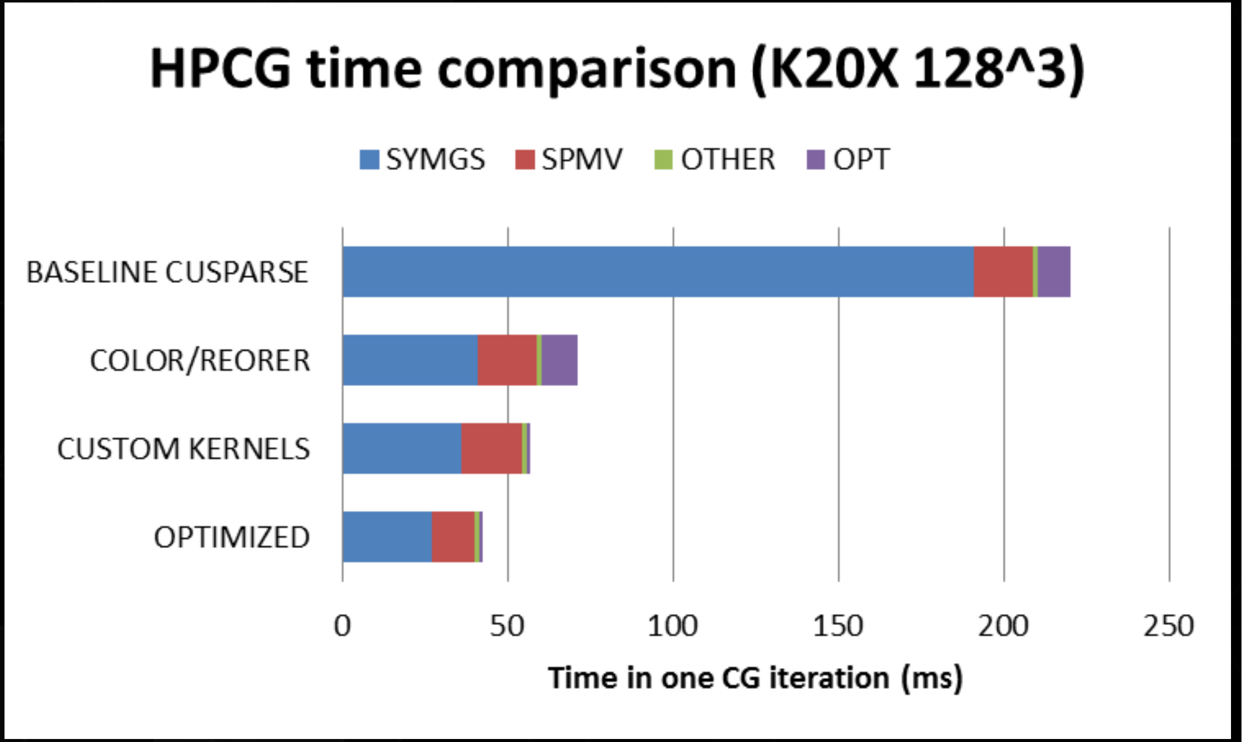
I. Cusparse CSR

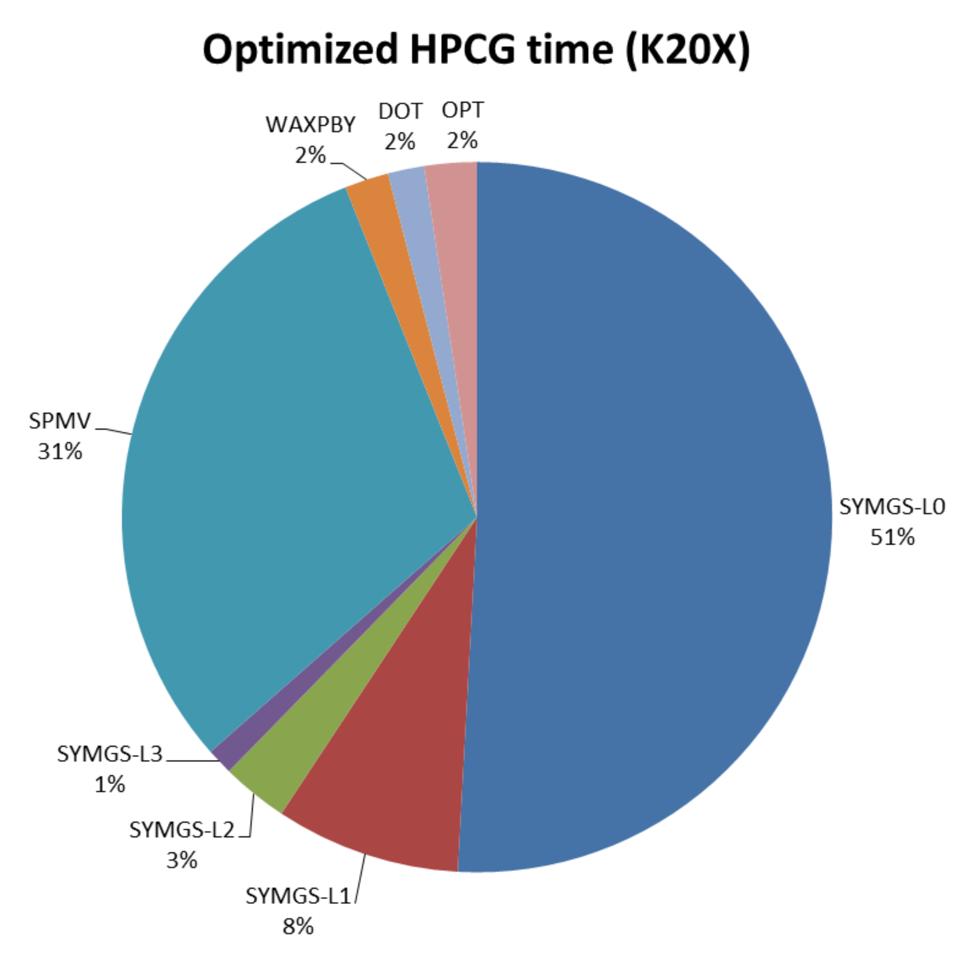
II. Cusparse CSR + Matrix Reordering (graph coloring)

III. Custom Kernels CSR + Matrix Reordering (graph coloring)

IV. Custom Kernels ELL + Matrix Reordering (graph coloring)

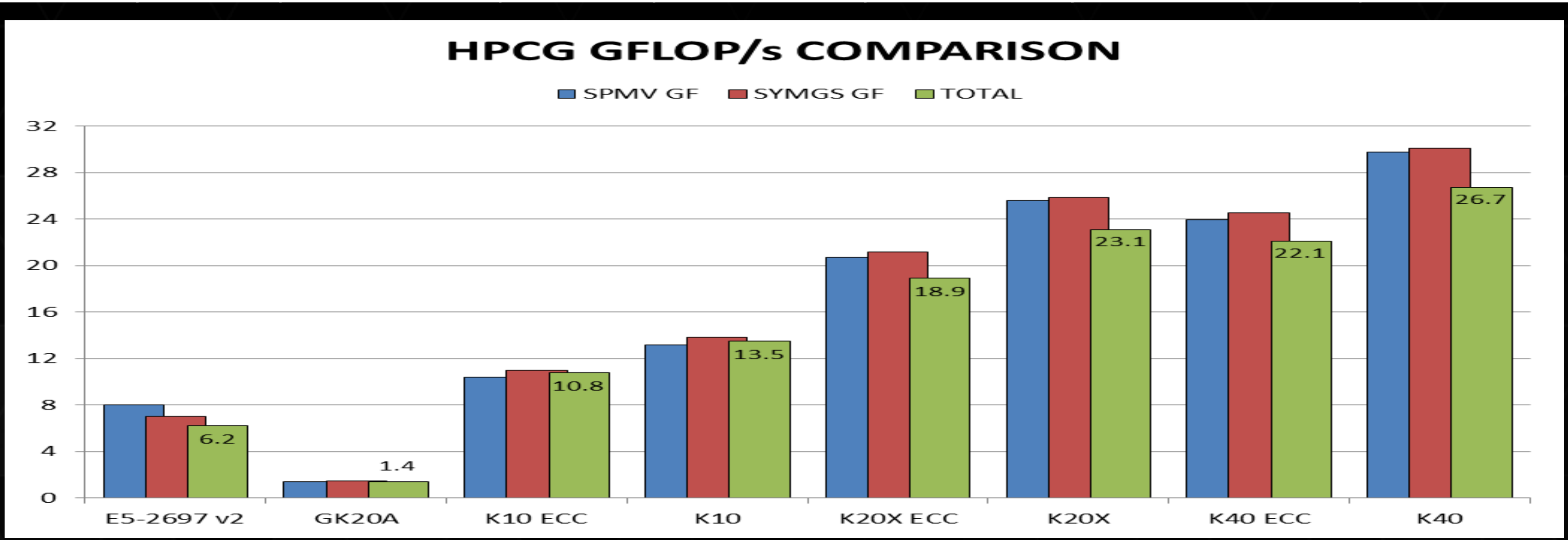
RESULTS – SINGLE GPU



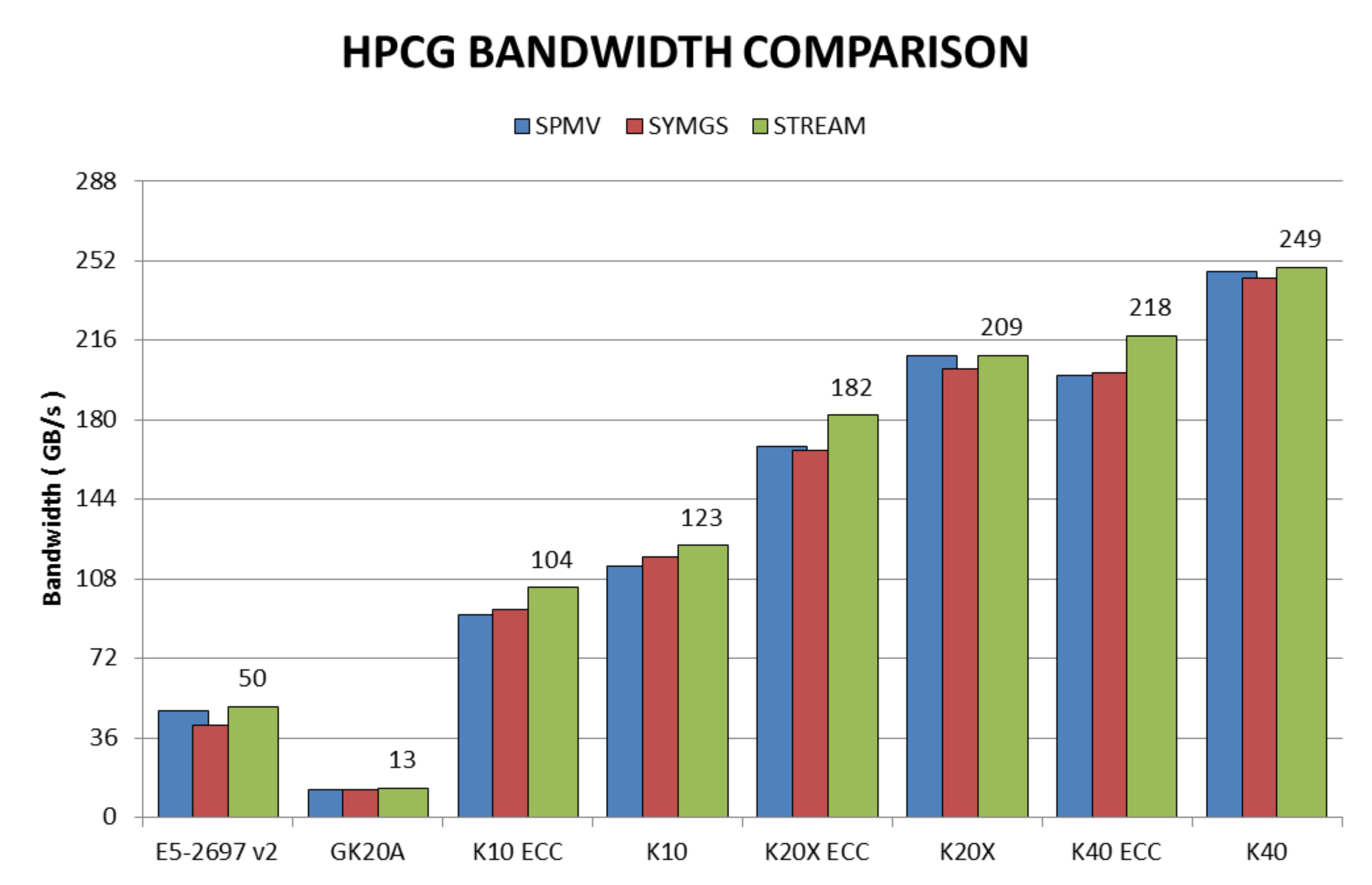


RESULTS – SINGLE GPU

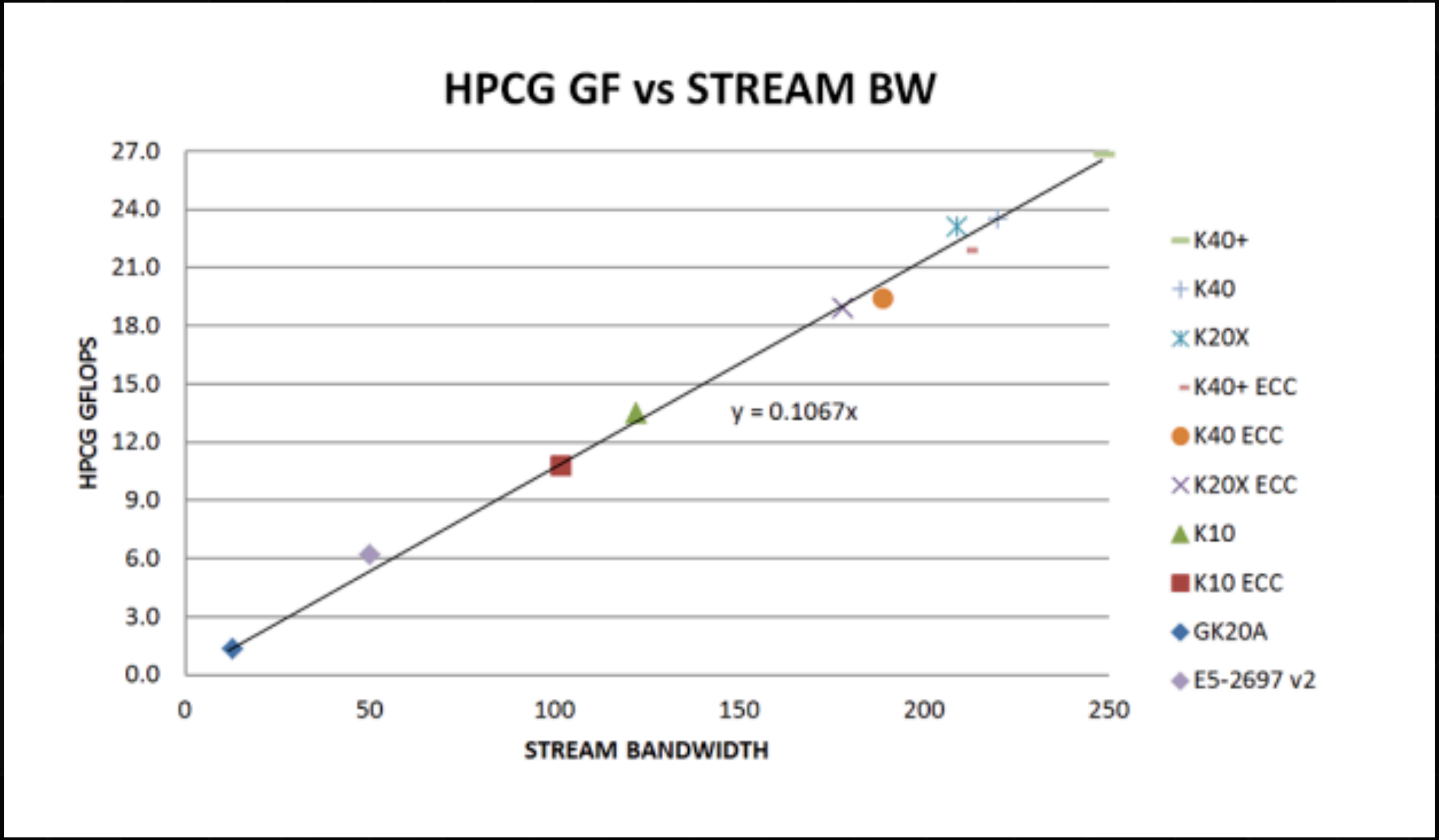
GPU	#SMs	#Cores SP/DP	Core Clock	DP (Gflops)	Memory Clock	Memory Bus Width	Memory Bandwidth
Tegra K1	1	192/8	852	13.6	924	64-bit	14.7
Tesla K10	8	1536/64	745	95	2500	256-bit	160
Tesla K20x	14	2688/896	732	1310	2600	384-bit	250
Tesla K40	15	2880/960	875	1680	3000	384-bit	288



RESULTS – SINGLE GPU

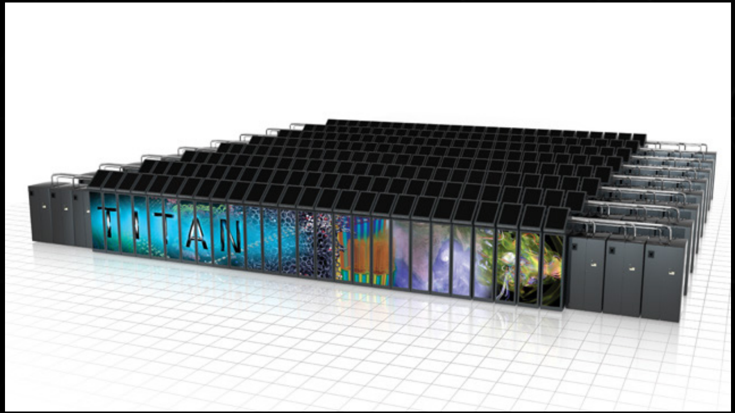


RESULTS – SINGLE GPU



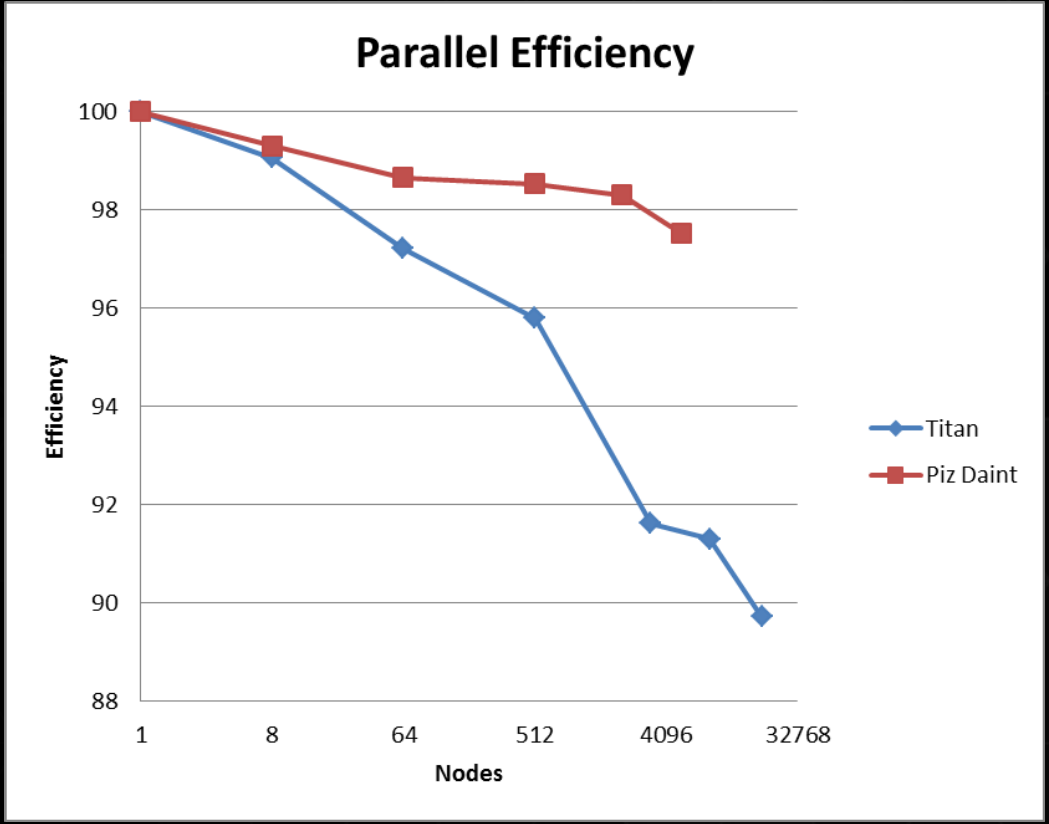
RESULTS – GPU SUPERCOMPUTERS

- Titan @ ORNL
  - Cray XK7, 18688 Nodes
  - 16-core AMD Interlagos + K20X
  - Gemini Network - 3D Torus Topology
- Piz Daint @ CSCS
  - Cray XC30, 5272 Nodes
  - 8-core Xeon E5 + K20X
  - Aries Network – Dragonfly Topology



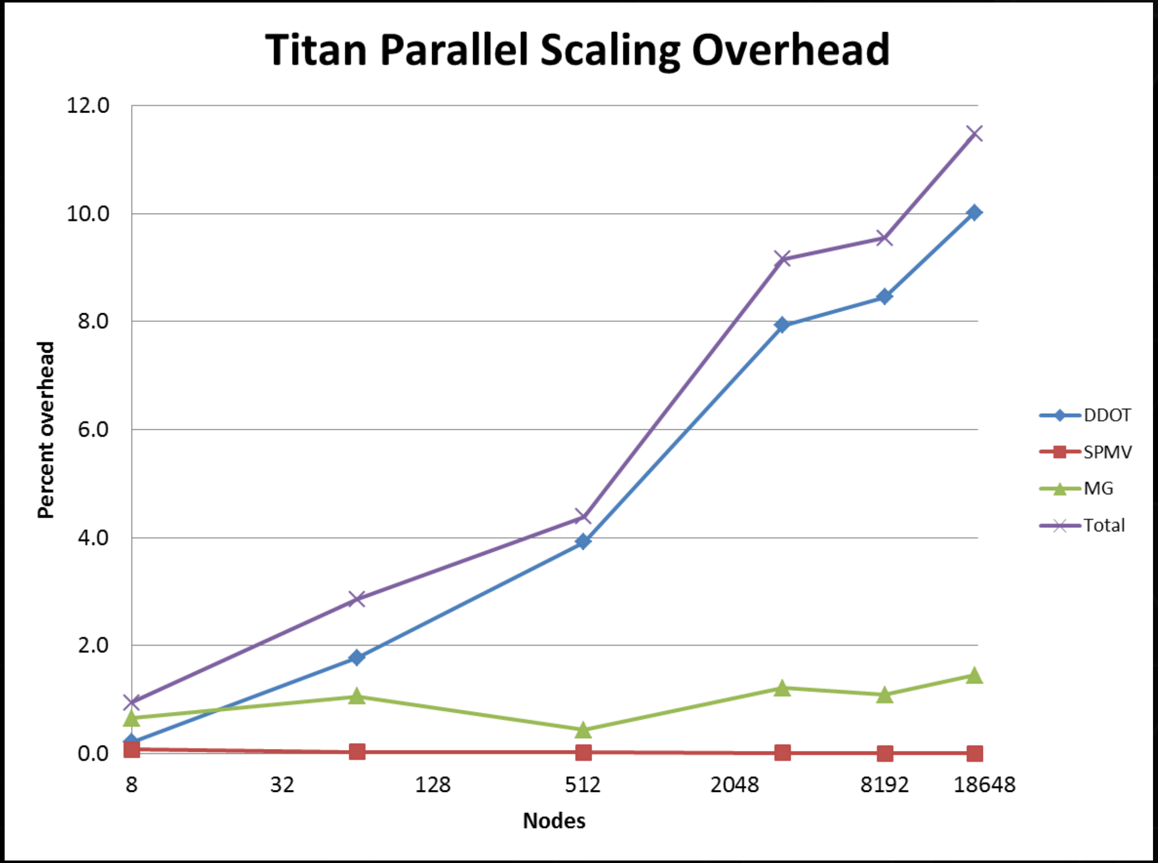
RESULTS – GPU SUPERCOMPUTERS

- 1 GPU = 20.8 GFLOPS (ECC ON)
- ~7% iteration overhead at scale
- Titan @ ORNL
  - 322 TFLOPS (18648 K20X)
  - 89% efficiency (17.3 GF per GPU)
- Piz Daint @ CSCS
  - 97 TFLOPS (5265 K20X)
  - 97% efficiency (19.0 GF per GPU)



RESULTS – GPU SUPERCOMPUTERS

- DDOT (-10%)
  - MPI_Allreduce()
  - Scales as Log(#nodes)
- MG (-2%)
  - Exchange Halo (neighbor)
- SPMV (-0%)
  - Overlapped w/Compute



REPRODUCIBILITY

- Residual Variance (reported in output file)
  - zero = deterministic order of floating point operations
- GPU Supercomputers bitwise reproducible up to full scale
  - except with network hardware-acceleration enabled on Cray XC30
- Parallel Dot Product
  - Local GPU routines bitwise reproducible
  - MPI_Allreduce()
    - reproducible with default MPI implementation
    - Non-reproducible with network offload (hardware atomics)

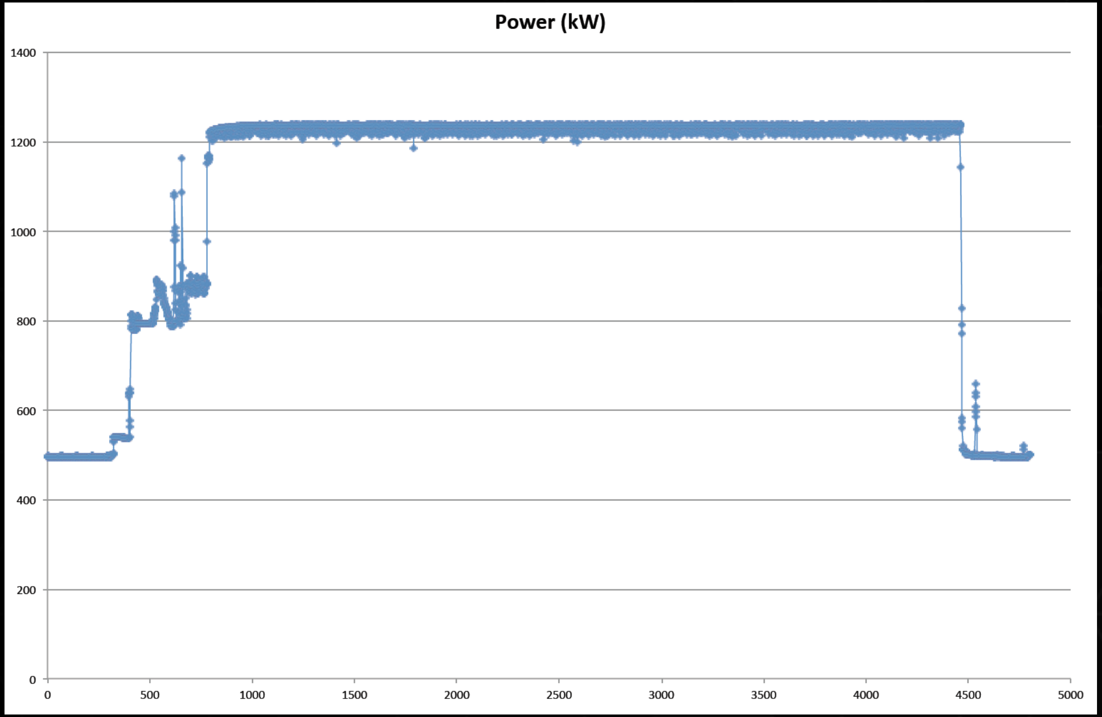
REPRODUCIBILITY

- CRAY XC30 MPI_Allreduce()
  - Default → reproducible results but lower performance
    - Min MPI_Allreduce time: 0.0296645
      Max MPI_Allreduce time: 0.153267
      Avg MPI_Allreduce time: 0.0916832
  - MPICH_USE_DMAPP_COL=1
    - Min DDOT MPI_Allreduce time: 0.0379143
      Max DDOT MPI_Allreduce time: 0.0379143
      Avg DDOT MPI_Allreduce time: 0.0379143
    - Residuals:

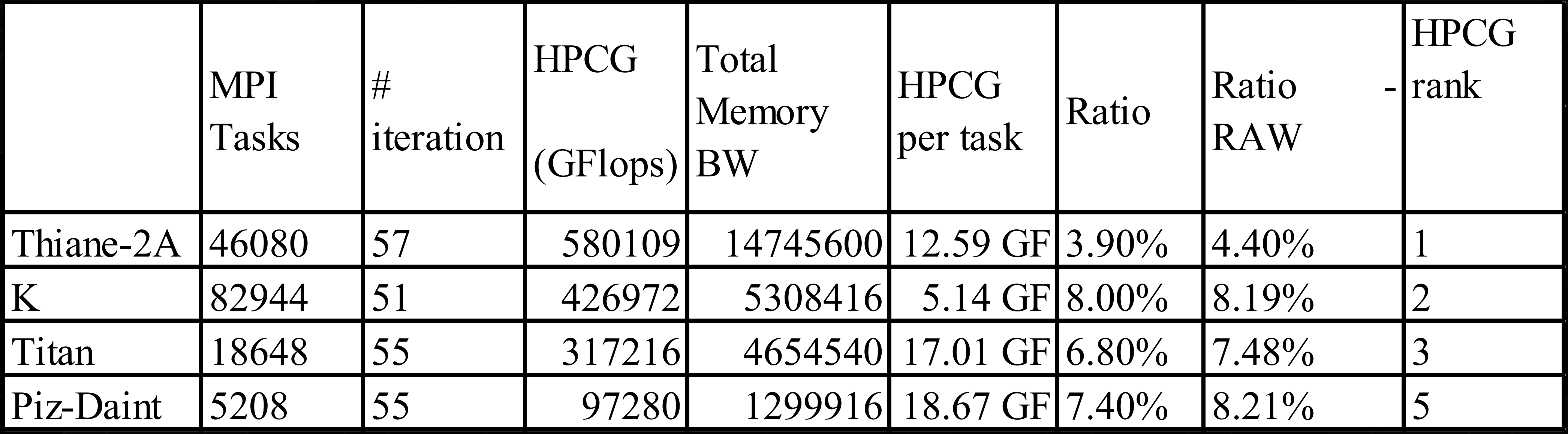
```
4.25079640861055e-08
4.25079640861032e-08
4.25079640861079e-08
4.25079640861054e-08
```

POWER CONSUMPTION

- Piz Daint (5208 K20X)
  - 99 TF / 1232 kW
  - 0.080 GF/W
- GK20A (Jetson TK1)
  - 1.4 GF / 8.3 Watts
  - 0.168 GF/W



PLATFORM COMPARISON



	MPI Tasks	# iteration	HPCG (GFlops)	Total Memory BW	HPCG per task	Ratio	Ratio - RAW	HPCG rank
Thiane-2A	46080	57	580109	14745600	12.59 GF	3.90%	4.40%	1
K	82944	51	426972	5308416	5.14 GF	8.00%	8.19%	2
Titan	18648	55	317216	4654540	17.01 GF	6.80%	7.48%	3
Piz-Daint	5208	55	97280	1299916	18.67 GF	7.40%	8.21%	5

Data from ISC14

CONCLUSIONS/ SUGGESTIONS

- (C) GPUs proven effective for HPL, especially for power efficiency
  - High flop rate
- (C) GPUs also very effective for HPCG
  - High memory bandwidth (Stacked memory will give a huge boost)
- (S) Reduce the required runtime from 1h to at least 100 iterations
- (S) Change metric: DOF/s?
- (S) Include yaml files in the list
- (S) Add power consumption?

ACKNOWLEDGMENTS

- Oak Ridge Leadership Computing Facility (ORNL)
  - Buddy Bland, Jack Wells and Don Maxwell
- Swiss National Supercomputing Center (CSCS)
  - Gilles Fourestey and Thomas Schulthess
- NVIDIA
  - Lung Scheng Chien and Jonathan Cohen