UPDATE ON THE CUDA IMPLEMENTATION OF THE HPCG BENCHMARK

Everett Phillips, Massimiliano Fatica

HPCG 3.0

- Optimized CUDA versions available from the official HPCG web site
- Support for latest CUDA (8.0) and latest hardware (M40, P100)
- Additional optimizations in the setup phase to improve performance on nodes with multiple GPUs:

	1 K80 (2 GPUs)	2 K80 (4 GPUs)	4 K80 (8 GPUs)
Previous version	55.2 GF (418.7 GB/s Effective) 27.6 GF_per (209.3 GB/s Effective)	109.2 GF (827.9 GB/s Effective) 27.3 GF_per (207.0 GB/s Effective)	204.4 GF (1549.7 GB/s Effective) 25.5 GF_per (193.7 GB/s Effective)
Latest version	55.9 GF (424.2 GB/s Effective) 28.0 GF_per (212.1 GB/s Effective)	111.3 GF (844.2 GB/s Effective) 27.8 GF_per (211.1 GB/s Effective)	211.0 GF (1600.4 GB/s Effective) 26.4 GF_per (200.0 GB/s Effective)

HPCG 3.0 RESULTS

```
1 x K80 (2 GPUs), ECC enabled, clk=875

2x1x1 process grid
256x256x256 local domain
SpMV  =  49.1 GF ( 309.1 GB/s Effective)  24.5 GF_per ( 154.6 GB/s Effective)
SymGS =  62.2 GF ( 480.2 GB/s Effective)  31.1 GF_per ( 240.1 GB/s Effective)
total =  58.7 GF ( 445.3 GB/s Effective)  29.4 GF_per ( 222.7 GB/s Effective)
final =  55.1 GF ( 417.5 GB/s Effective)  27.5 GF_per ( 208.8 GB/s Effective)
```

1.8% of peak

```
2 x M40, ECC enabled, clk=1114

2x1x1 process grid
256x256x256 local domain
SpMV  =  69.4 GF ( 437.2 GB/s Effective)  34.7 GF_per ( 218.6 GB/s Effective)
SymGS =  83.7 GF ( 645.7 GB/s Effective)  41.8 GF_per ( 322.8 GB/s Effective)
total =  79.6 GF ( 603.7 GB/s Effective)  39.8 GF_per ( 301.9 GB/s Effective)
final =  74.2 GF ( 562.7 GB/s Effective)  37.1 GF_per ( 281.4 GB/s Effective)
```

18% of peak

```
1 x K80 (2 GPUs), ECC disabled, clk=875

2x1x1 process grid
256x256x256 local domain
SpMV  =  59.9 GF ( 377.0 GB/s Effective)  29.9 GF_per ( 188.5 GB/s Effective)
SymGS =  74.6 GF ( 575.8 GB/s Effective)  37.3 GF_per ( 287.9 GB/s Effective)
total =  70.6 GF ( 535.5 GB/s Effective)  35.3 GF_per ( 267.7 GB/s Effective)
final =  66.0 GF ( 500.2 GB/s Effective)  33.0 GF_per ( 250.1 GB/s Effective)
```

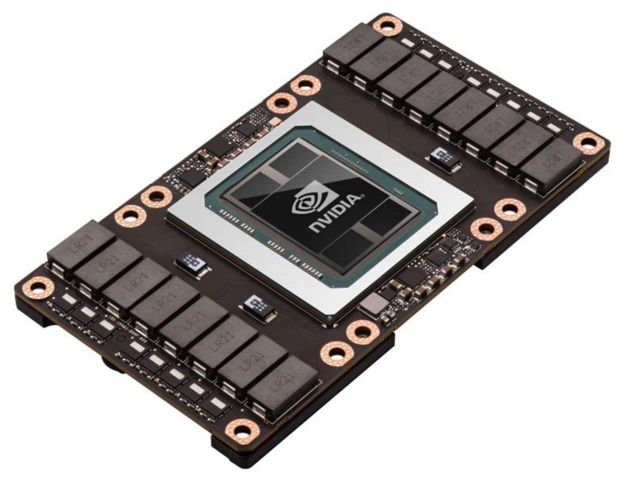
```
2 x M40, ECC disabled, clk=1114

2x1x1 process grid
256x256x256 local domain
SpMV  =  79.0 GF ( 497.9 GB/s Effective)  39.5 GF_per ( 248.9 GB/s Effective)
SymGS =  95.9 GF ( 740.5 GB/s Effective)  48.0 GF_per ( 370.2 GB/s Effective)
total =  91.1 GF ( 691.1 GB/s Effective)  45.6 GF_per ( 345.6 GB/s Effective)
final =  84.9 GF ( 644.2 GB/s Effective)  42.5 GF_per ( 322.1 GB/s Effective)
```

M40: GM200 chip, DP:SP=1:32, SP=7TF, DP=0.21TF, 384bit at 6 GHz

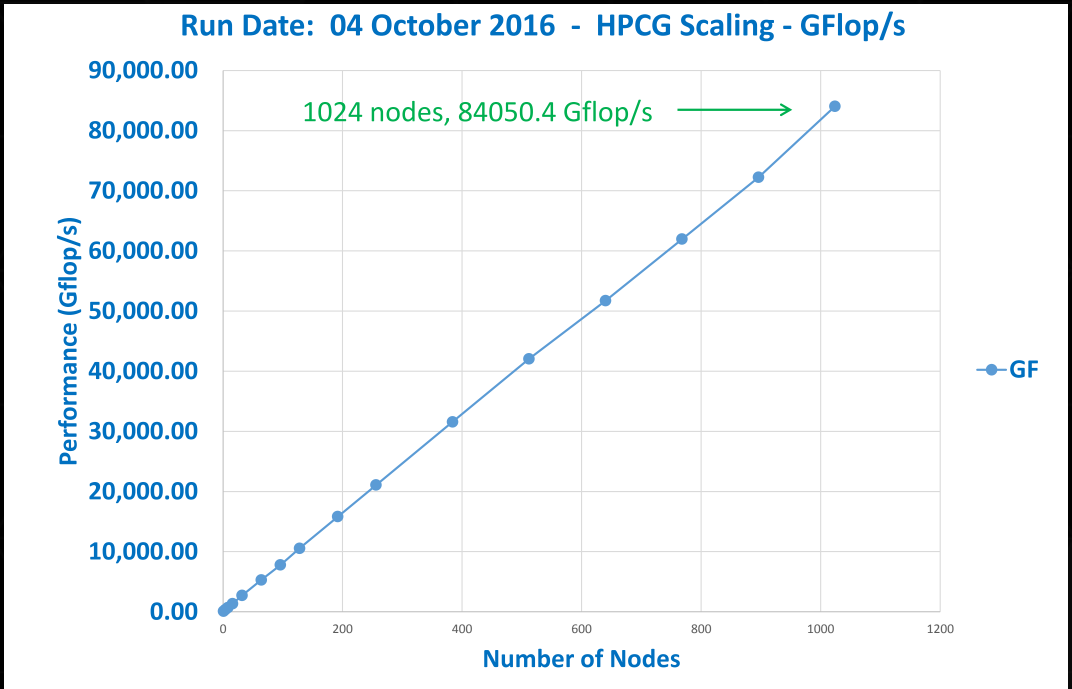
K80 : 2 xGK210 , DP:SP=1:3, SP=8.74TF, DP=2.91TF, 2x384 bit at 5GHz

HPCG ON PASCAL

- HPCG is all about memory bandwidth
- TESLA P100 is the first GPU with HBM2 memory:
  - Two form factors, PCI-e and SXM2, have different core clocks but same memory clock:
    - 715 MHz, 4096 bit memory controller:
      - 715 GB/s peak, ~520 GB/s STREAM
    - No performance penalty for ECC
  - Fastest processor to run HPCG (>80 GF per GPU)



PIZ DAINT SCALING



PIZ DAINT RESULTS

```
HPCG-Benchmark
3.0
Release date: November 11, 2015
Machine Summary:
  Distributed Processes: 3024
  Threads per processes: 12
Global Problem Dimensions:
  Global nx: 3072
  Global ny: 4608
  Global nz: 3584
Processor Dimensions:
  npx: 12
  npy: 18
  npz: 14
Local Domain Dimensions:
  nx: 256
  ny: 256
  nz: 256
__________ Final Summary __________:
  HPCG result is VALID with a GFLOP/s rating of: 247981
      HPCG 2.4 Rating (for historical value) is: 256984
```

82 GF per node

(#8 on current list)