GPUs and NVIDIA Q1 results

Economic drivers

Graphics Co-processors

- Early graphics gaming systems (pre-1990s) in mass market products were typically 2D
  - Hardware concerned with rasterizing, object collisions, etc.
- Rise of 3D games (with ray-tracing, sophisticated rendering, etc) and home consoles (or PC) drove a new paradigm: Graphics as a Linear Algebra Accelerator.
  - Problems of moving or rotating modeled objects are transforms
  - Problems of determining direction of reflected light, etc.

Linear Operations

- Often organized into “layers” of library calls:
  - Lowest layer: Vector primitives AXPY (A [times]X plus Y) “axe-pee”, scale, DOT (dot product)
  - Middle layer: Matrix multiplication of vector (or matrix)
  - Top Layer: LU decomposition (i.e. solve system), Invert Matrix, etc.
  - Reference: “BLAS” https://www.gnu.org/software/gsl/manual/html_node/BLAS-Support.html

How to speed up?

- Add instructions to CPU?
  - Intel MMX (c2000) has load/store & arithmetic on chunks of vectors
  - Might handle fixed length only (but would accelerate libraries- can partially “unroll” code loops)
  - May impact cacheing- “striding” through memory can defeat cache
- Make a separate GPU?
  - Can also have own memory interface for frame buffer, with different (optimized) memory width/caching
  - ATI (Now AMD), NVIDIA (Mid 1980s – Mid 1990s)

But what else can we do with it?

- Many computational problems can be cast as linear algebra
- Graphics cards can also address “embarrassingly parallel” computation, even if not vector/matrix operations
- 2007: Nvidia releases CUDA API for GPUs
  - Linear algebra libraries
  - FFT (signal processing) libraries
  - Etc.
- It’s not just for graphics anymore…
  - Problems formerly in the realm of national supercomputer centers can now be approached with modest-sized servers
- 2017: Nvidia V100 adds support for direct 4x4 matrix multiplication

Example: Intel I7 vs Nvidia GTX 1080

(Chrzeszczyk & Anders, 2017 https://developer.nvidia.com/sites/default/files/akamai/cuda/files/Misc/mygpu.pdf)

Solving the general NxN linear system in single precision.

./testing_sgesv --lapack

```
%   N  NRHS   CPU Gflop/s (sec)   GPU Gflop/s (sec)
%=============================================
1088    1    75.89 (   0.01)    74.72 (   0.01)
9280    1   285.11 (   1.87)   1220.60 (   0.44)
```

Matrix-matrix product in single precision.

./testing_sgemm --lapack

```
% transA = No transpose, transB = No transpose
%   M     N     K   MAGMA Gflop/s (ms)  cuBLAS Gflop/s (ms)   CPU Gflop/s (ms)
%=================================================================================
 1088  1088  1088     2985.30 (   0.86)     5377.70 (   0.48)     344.92 (   7.47)
10304 10304 10304   4820.21 ( 453.92)    7505.78 ( 291.51)    418.97 (5222.33)
```

Nvidia market segments Nvidia Q1 results, subset

• Gaming revenue grew 68 percent from a year earlier to $1.72 billion.

• Datacenter revenue grew 71 percent from a year earlier to a record $701 million.
• Announced TensorRT 4™, the latest version of the TensorRT AI inference accelerator software, expanding its reach in the inference market by accelerating deep learning across a much broader range of applications.
• Announced GPU acceleration for Kubernetes to facilitate enterprise inference deployment on multi-cloud GPU clusters.

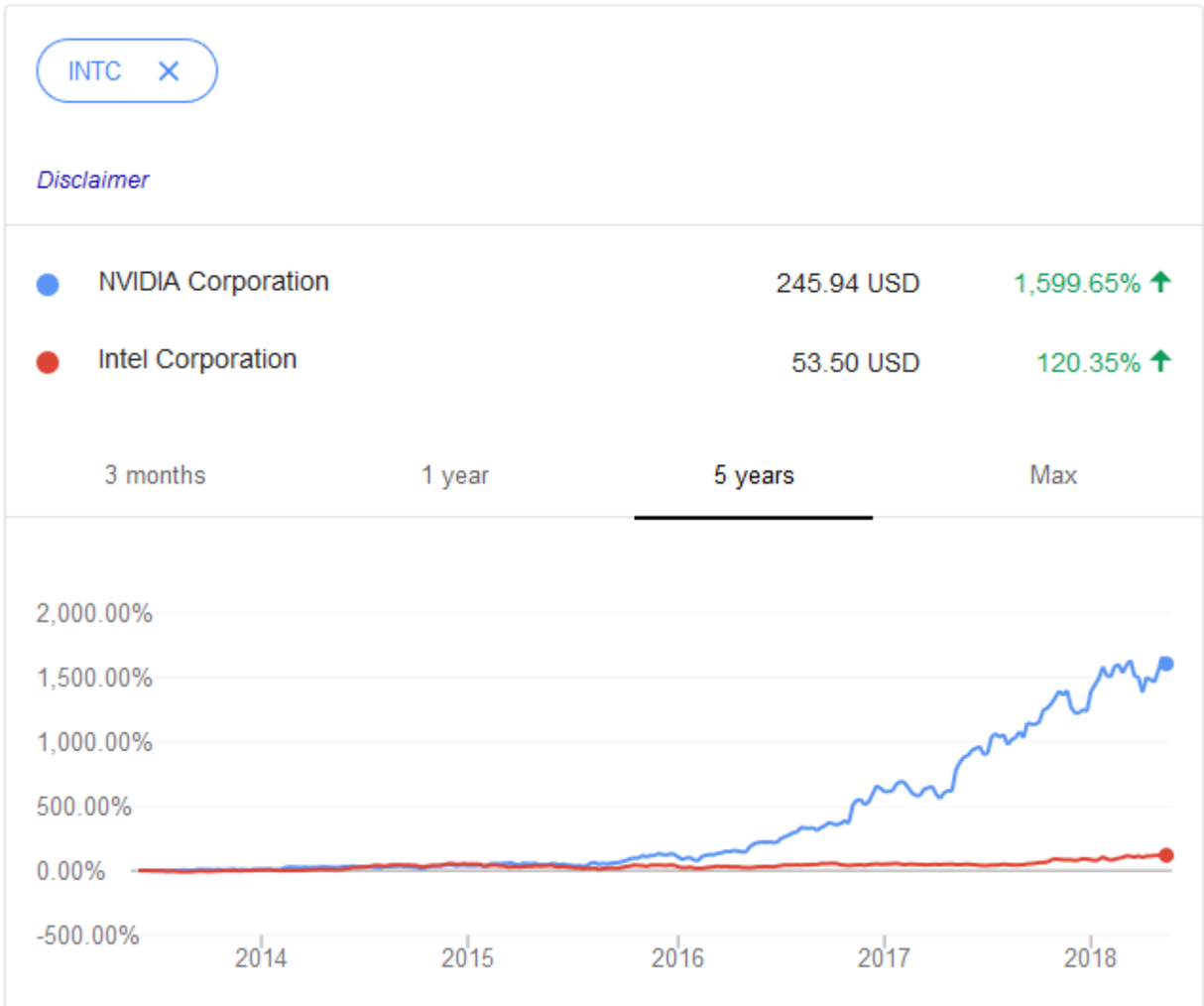
• Professional Visualization revenue grew 22 percent from a year earlier to $251 million.
• Announced the Quadro® GV100 GPU with RTX technology, making real-time ray tracing possible on professional design and content creation applications.

• Automotive revenue grew 4 percent from a year earlier to a record $145 million.

New Platforms
• Introduced Project Clara, a medical imaging supercomputer, to revolutionize medical imaging.
• Announced that Arm will integrate the open-source NVIDIA Deep Learning Accelerator to bring AI inference to mobile, consumer electronics and Internet of Things devices.

Nvidia stock price (From Google Finance)

Nvidia & Intel stock price (From Google Finance)



Nvidia challenge & strategy

- How to manage price discrimination?
  - i.e. How to charge each customer's maximum willingness to pay?
- But Gaming is currently biggest part of business
  - And if they raise prices there, they invite competition at low end (which could eventually compete at high end)
  - Don't want situation like CPUs and HDDs, where cloud providers moved to lowest cost commodity HW
- And Cryptocurrency (i.e. "Proof of effort computations") is altering demand
    - But also may be helping the PE multiple of their stock….

Answer(?): A “non-market” strategy

- Set up a legal barrier to prevent any “deep pocket” datacenter operators from using gaming cards.
  - Might not actively pursue “small fry,” but would likely go after Fortune 100 and major clouds
- But…. Don’t want to upset the analysts by deterring cryptocurrency boom (bubble?), so; this curious license on accelerator cards:

“No Datacenter Deployment. The SOFTWARE is not licensed for datacenter deployment, except that blockchain processing in a datacenter is permitted.” (So no licenses or libraries, etc.)

Likely near term impact:

- Nvidia likely maintains higher prices/margins near term
  - Might provide "price umbrella" giving other players (e.g. Intel, Xilinx) more room for a product.
- The "Big Money" cloud players (AWS, IBM, Google) likely develop their own hardware.
  - They also have big enough patent portfolios to defend against a suit.
  - Matrix Multiplication – probably hard to get a basic utility patent, but may be many patent-able aspects relating to specialized pipelining and cacheing (fast matrix hardware ineffective if memory access is inefficient)
- POSSIBLY reduces appeal in FOSS/academic community, with long-term impact
- POSSIBLY improves Nvidia near-term profitability at a risk of future market share.

Discussion

- Other interesting financials at:
  - http://financials.morningstar.com/ratios/r.html?t=INTC
  - http://financials.morningstar.com/ratios/r.html?t=NVDA