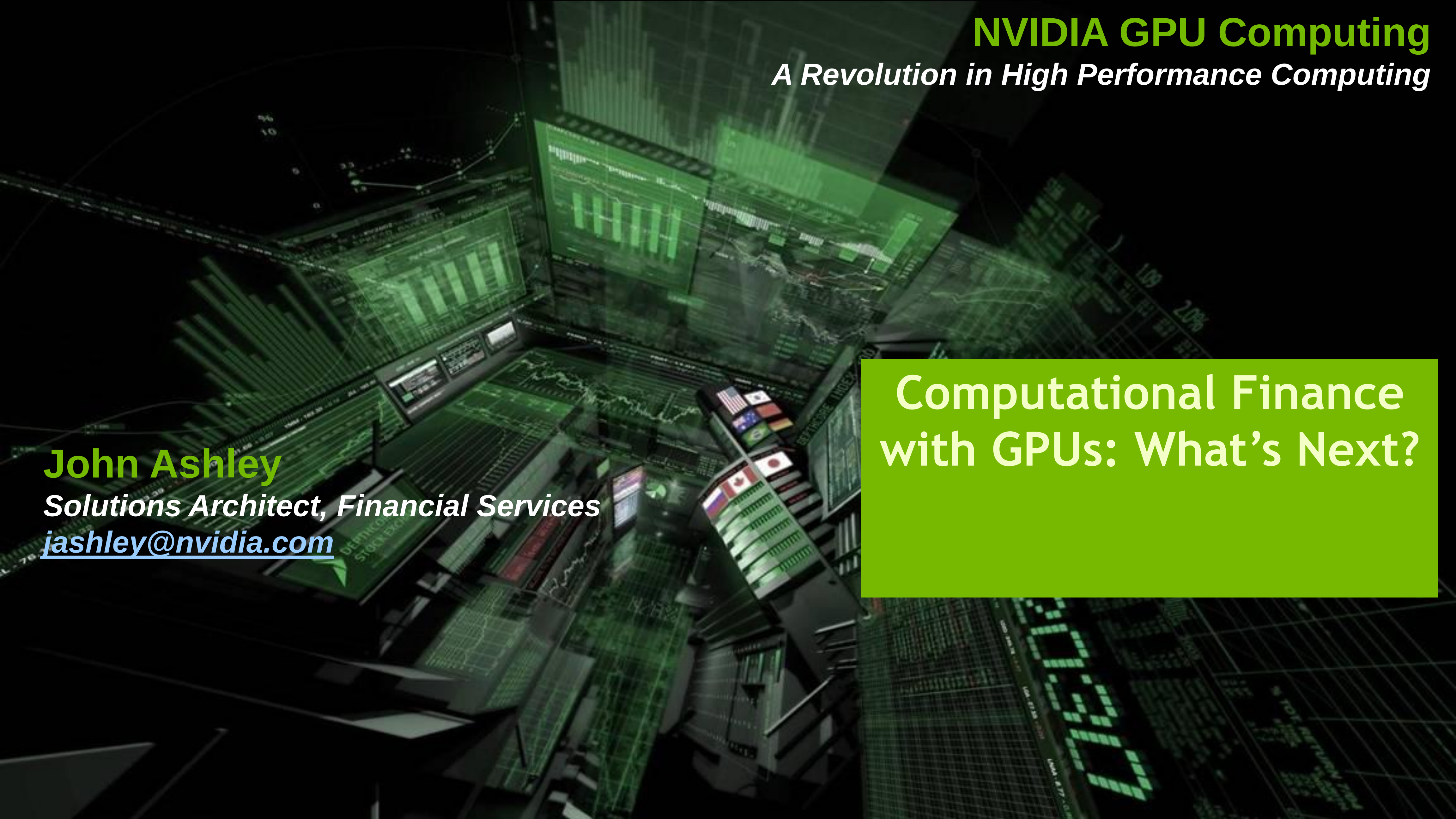


Computational Finance with GPUs: What's Next?

John Ashley

Solutions Architect, Financial Services

jashley@nvidia.com

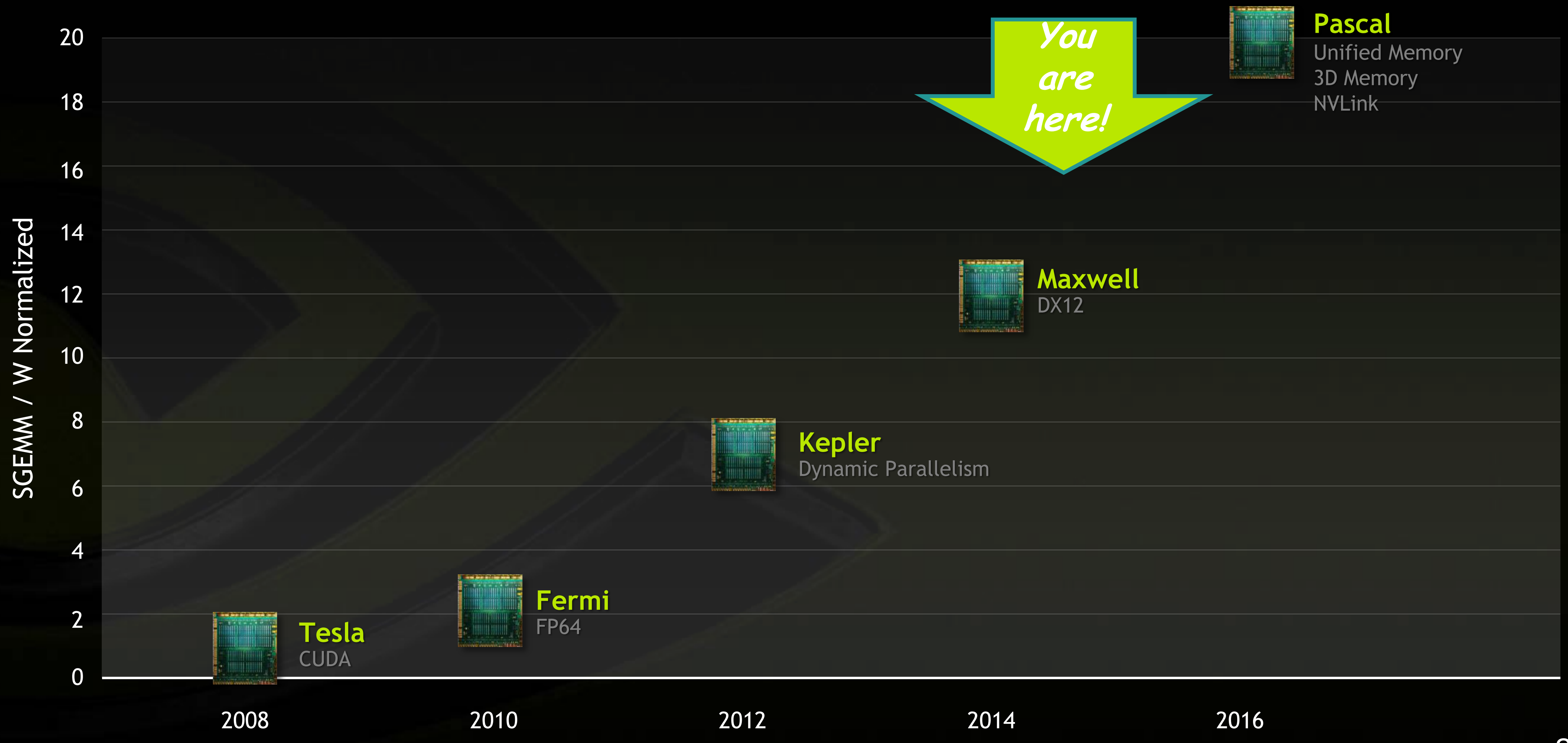
Computational Finance with GPUs: What's Next?

- Where have we come from?

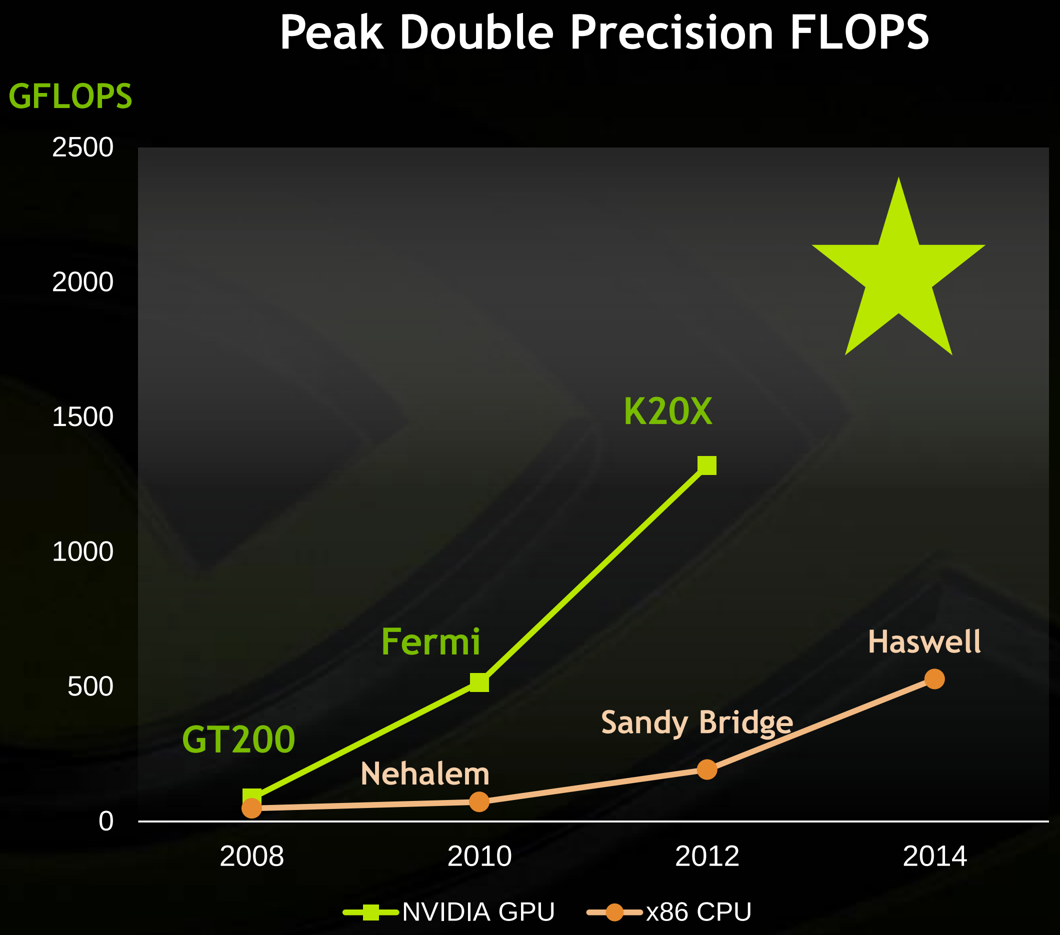
- Where are we now?

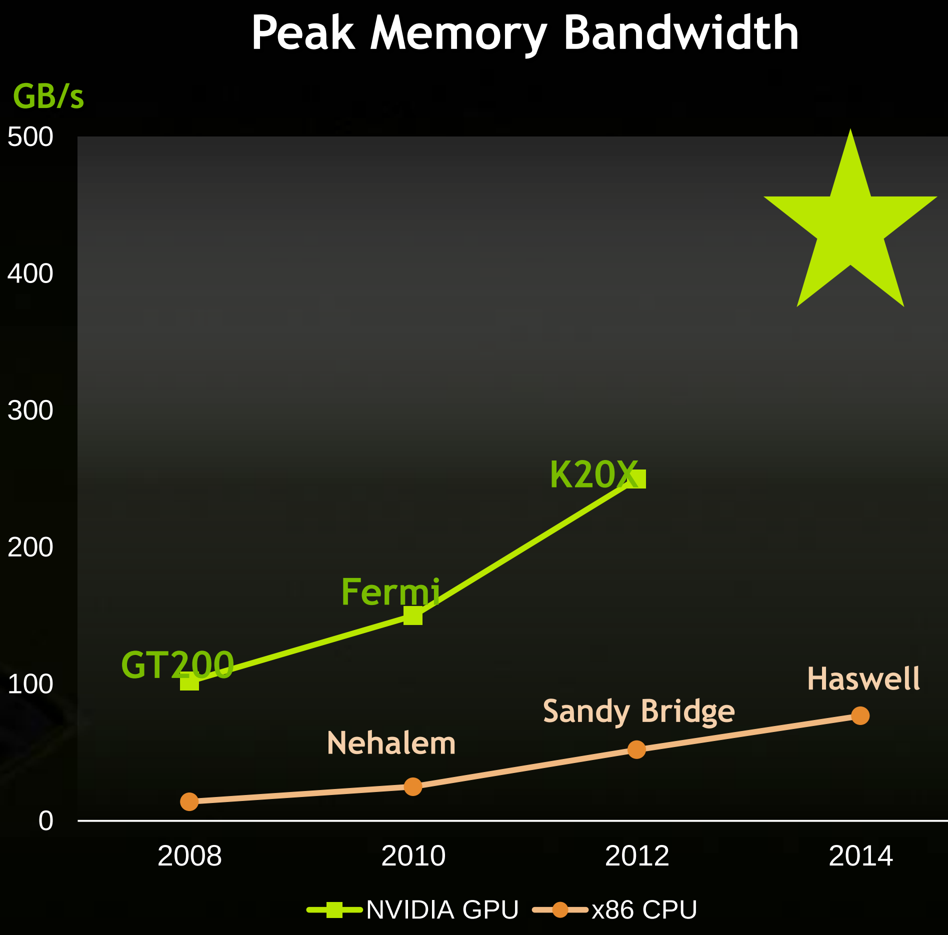
- Where are we going to?

Strong CUDA GPU Roadmap

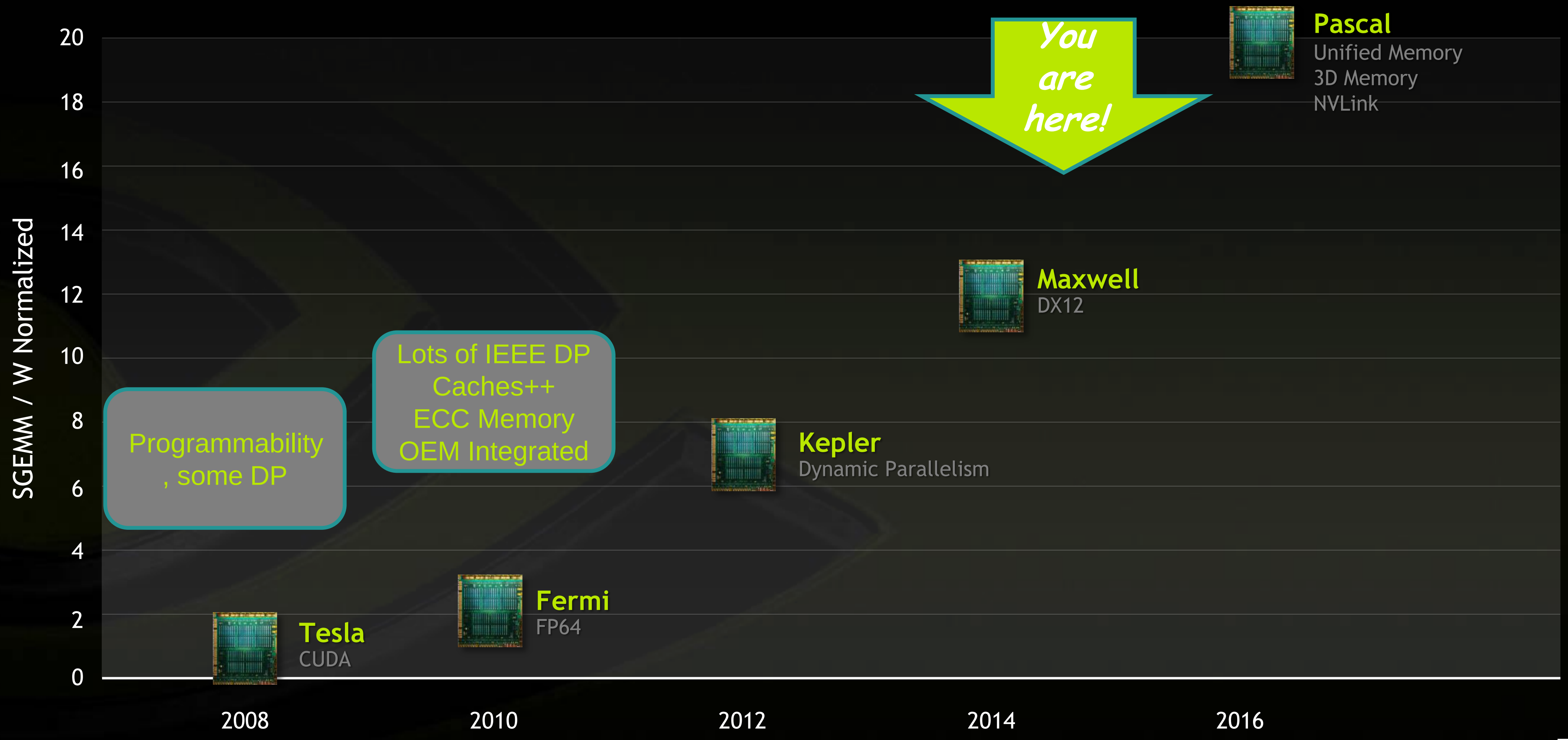


Performance Gap Continues to Grow





GPU Card Feature History



Programmability, some DP

Lots of IEEE DP
Caches++
ECC Memory
OEM Integrated

You are here!

Tesla
CUDA

Fermi
FP64

Kepler
Dynamic Parallelism

Maxwell
DX12

Pascal
Unified Memory
3D Memory
NVLink

Where have we come from? [Technology]

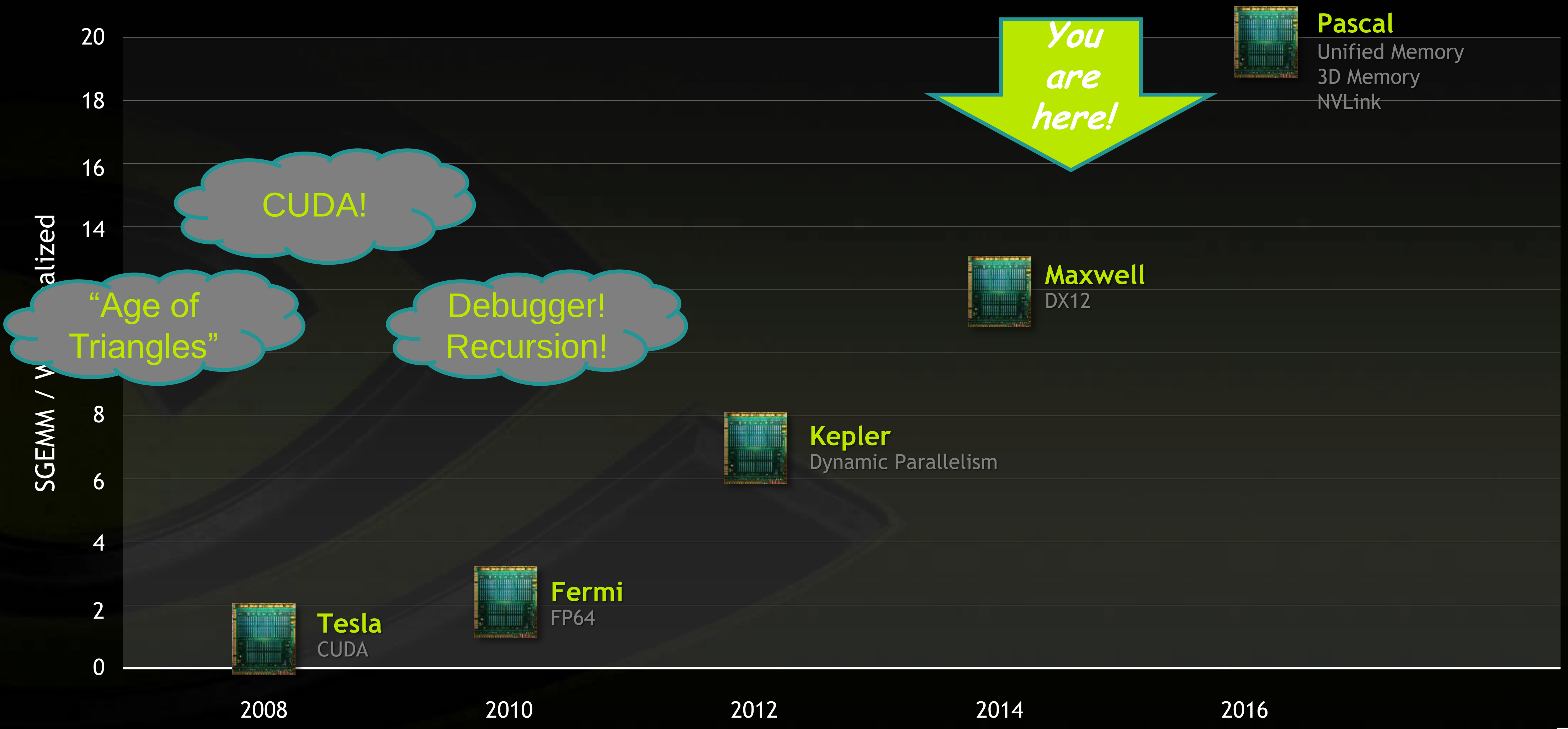
- The dark ages of GPU computing… before CUDA there was only OpenGL and shader languages – “programing with triangles”.

- Pre-2008 -- Before the S1070 (Tesla “Tesla”) GPUs had no double precision.

- S1070 / C1060 brought CUDA C++ and double precision support.
  - 240 cores, 4GB RAM, 933 GFLOPS SP, 77 GFLOPS DP, 102 GB/s
  - Aftermarket or custom build

- 2010 -- Fermi C/M 20xx – more DP, more BW, ECC, OEM Integrated…
  - Up to 512 cores, 6GB RAM
  - CUDA: Real function calls + Recursion

GPU “Programming History”



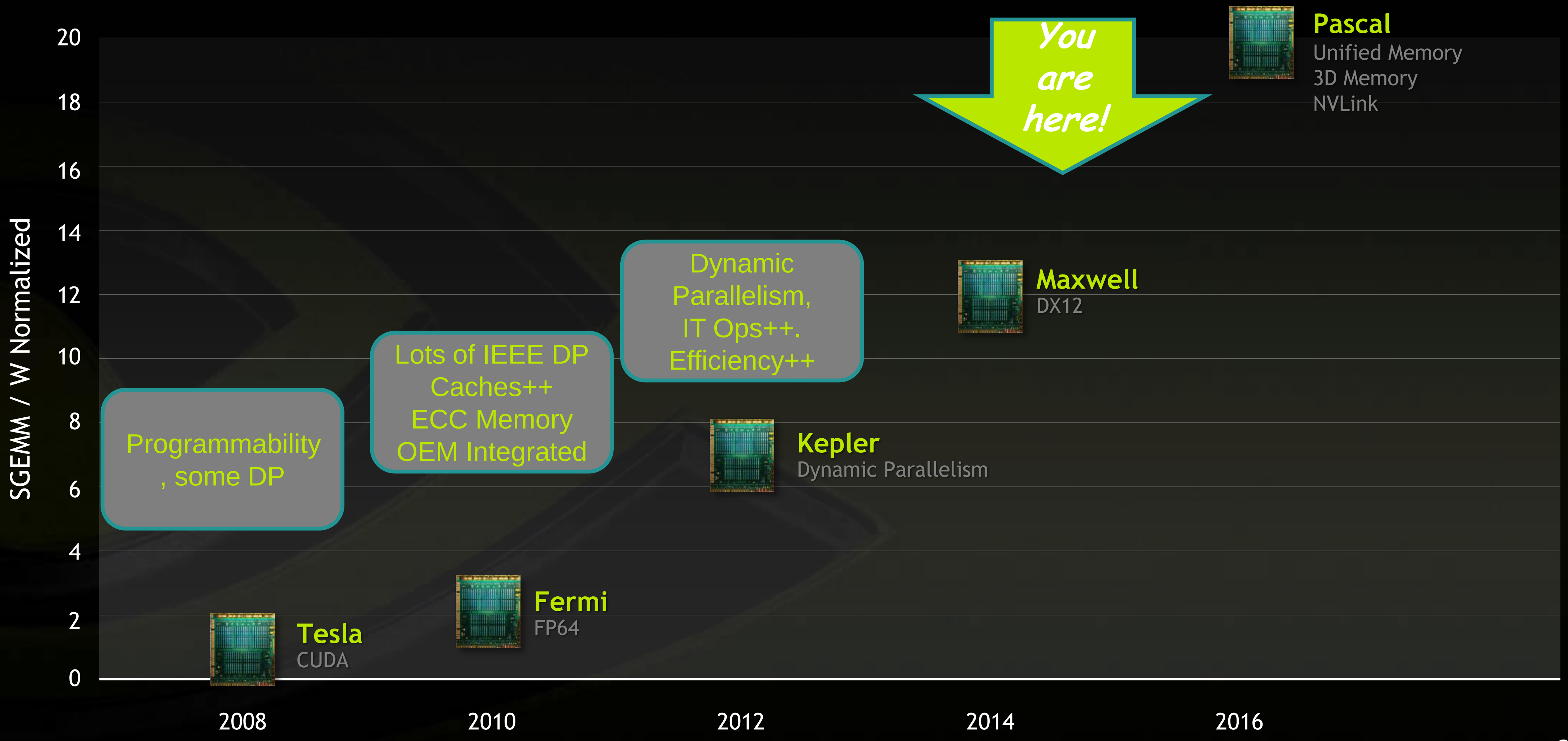
Where have we come from? [Finance]

- Pre-Fermi -- Pricing & Calibration
  - Early public use cases from Bloomberg & BNP Paribas
  - ISVs like Hanweck Associates, NAG

- Fermi brought the revolution
  - Additional DP perf and debuggers lead to easier programming
  - Still pricing, but also VaR models
  - Easier IT adoption via vendor supplied systems
  - First “business as usual” systems at banks
  - Insurance – Variable Annuity Hedging
  - Press releases by JPMC, Credit Agricole, others
  - ISVs like Murex, AON Benfield, Matlab, Altimesh, Xcelerit, SciComp, Mathematica, …

- Global Derivatives 2012
  - 2012 “Running Risk on GPUs”, D. Kandhai, ING Bank
  - 2012 “Combining Numerical & Technological Advances for Fast & Robust Monte Carlo Model Calibration”, J. Mahrun, Unicredit

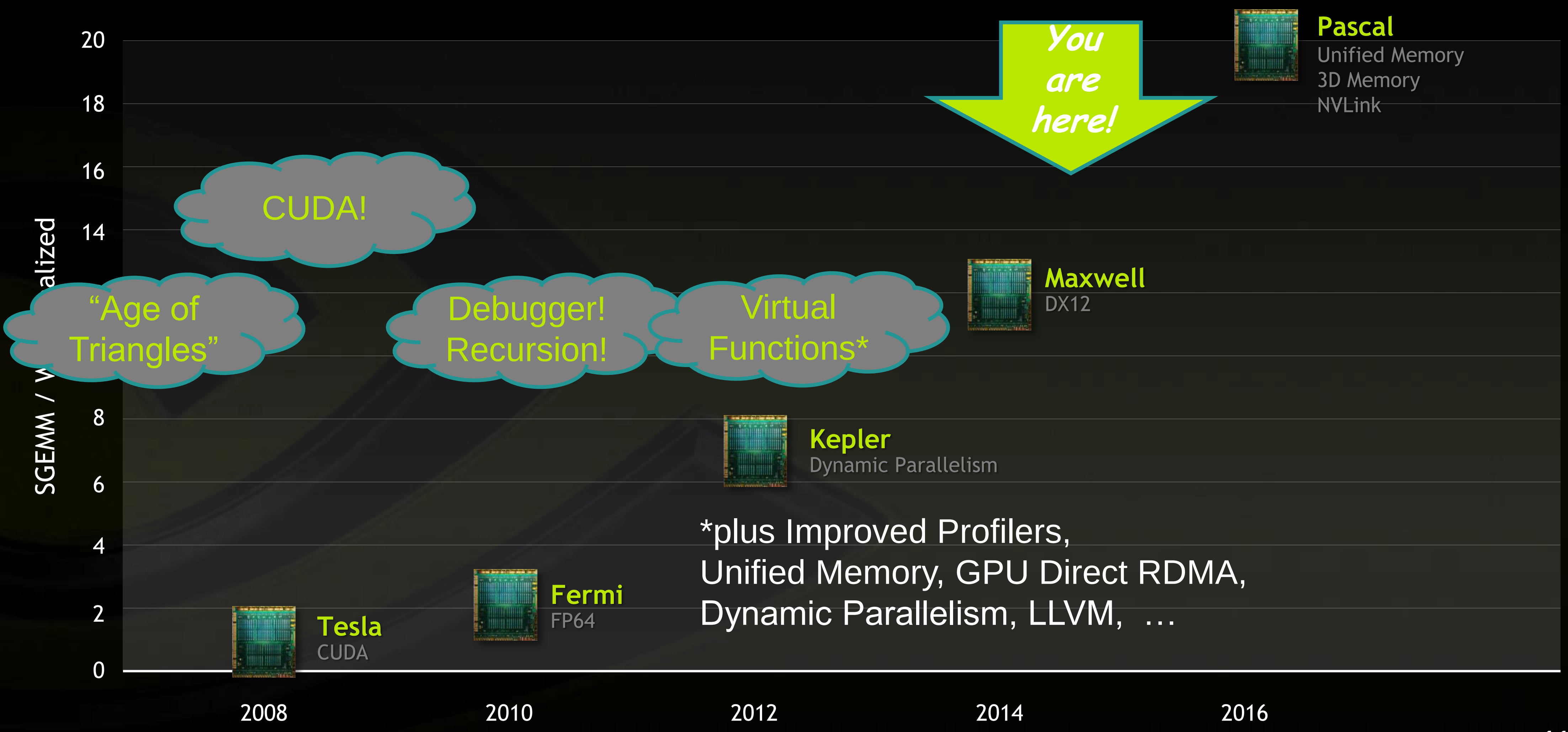
GPU Card Features Today



Where are we now? [Technology]

- Kepler K10
  - First compute GPU optimized for Single Precision performance
  - 2xGPU per card for higher density and better power efficiency
- Kepler K20/20x/40
  - 2496-2880 CUDA cores, 5-12 GB RAM, up to 288 GB/s, up to 1.4 DP TF
- CUDA
  - + Virtual Functions
  - + Dynamic Parallelism
  - + Improvements in debugging and profiling
- Language Partners
  - C#, F#, Python…

GPU “Programming History”



*plus Improved Profilers, Unified Memory, GPU Direct RDMA, Dynamic Parallelism, LLVM, …

Where are we now? [Finance]

- Biggest business driver is regulatory and business demand for CVA/DVA and especially FVA/Margining
- Cost reduction for overnight line of business risk
- Real time risk – better models, intra-day
- Even more ISVs
  - MiSys, QuantAlea, Sungard, MIMOS, Synerscope, Fuzzy Logic, …

Where are we now? [Finance]

- Global Derivatives 2013-2014
  - 2013 “From Parallel Algorithms To Monads: New Techniques For Using GPUs To Make Derivative Pricing & Risk Analysis More Efficient”, D. Egloff, QuantAlea
  - 2013 “GPU Acceleration for Interest Rate Modelling in Practice”, H. Wang, Barclays
  - 2014 “Leveraging GPU Technology For The Risk Management Of Interest Rates Derivatives”, G. Blacher and R. Smith, Bank of America Merrill Lynch
  - 2014 “Why GPU Tolls The Bell Of Gigantic CPU Grids For All Computation Intensive Use Cases Of The New Normal”, L. T. Nessi, Murex
- 5th Workshop on High Performance Computational Finance (WHPCF 2013)
- Computation in Finance and Insurance, post-Napier (Napier 400)
- University of Chicago “Recent Developments in Parallel Computing in Finance”

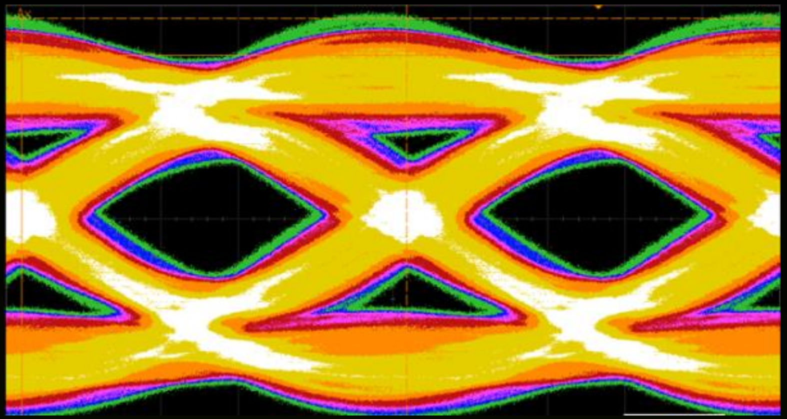
Where are we now? [Finance]

- GPU Technology Conference 2014/13
  - “Monte Carlo Simulation of American Options with GPUs”, J. Demouth, NVIDIA
  - “Effortless GPU Models for Finance”, B. Young, Sungard
  - “GPU Implementation of Explicit and Implicit Finite Difference Methods in Finance”, M. Giles, Oxford
  - “Accelerating Option Risk Analytics in R using GPUs”, M. Dixon, U. San Francisco
  - “GPU Enabled Real-time Risk Pricing in Option Market Marking”, C. Doloc, Chicago Trading Company
  - “High Performance Counterparty Risk and CVA Calculations in Risk Management”, D. Delarue and A. Siddiqi, BNP Paribas
  - “Domain Specific Languages for Financial Payoffs”, M. Leslie, Bank of America Merrill Lynch
  - “Hedge Strategy Simulation and Backtesting with DSLs, GPUs, and the Cloud”, A. Mohammad, Aon Benfield Securities

Where are we going? [Technology]

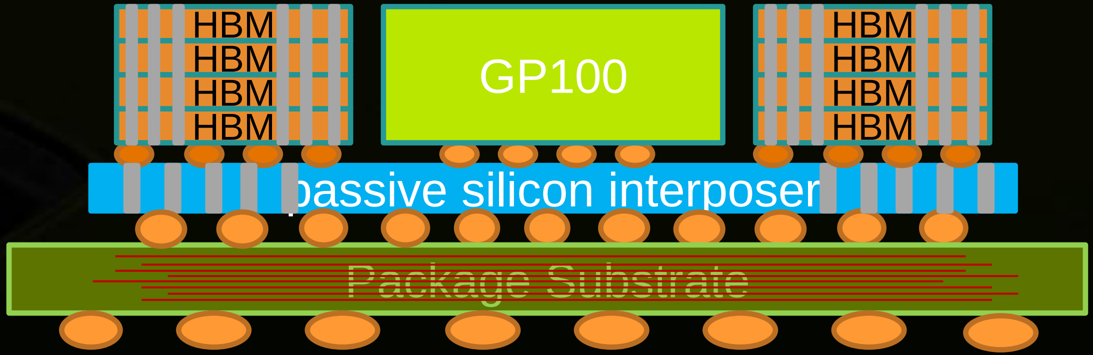
NVLINK

- GPU high speed interconnect
- 5-12x PCIe Gen 3 Bandwidth
- Drastically reduced energy/bit



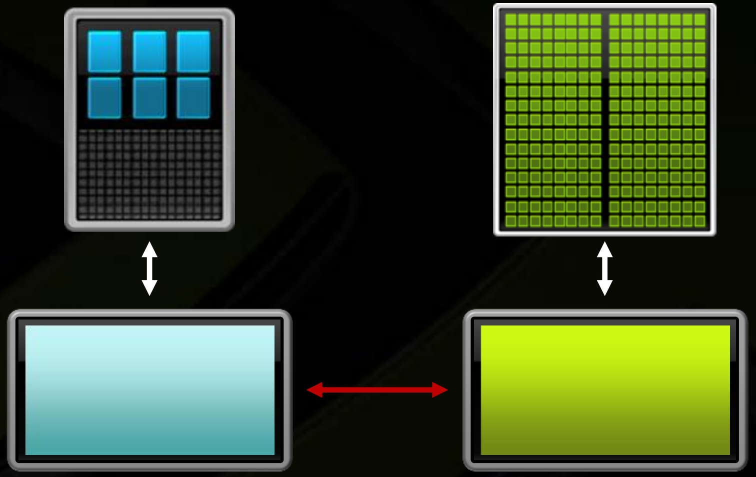
Stacked Memory

- 2-4x Capacity & Bandwidth
- 3-4x More Energy Efficient per bit
- Leaves more power for compute



Unified Memory -- Lower Developer Effort

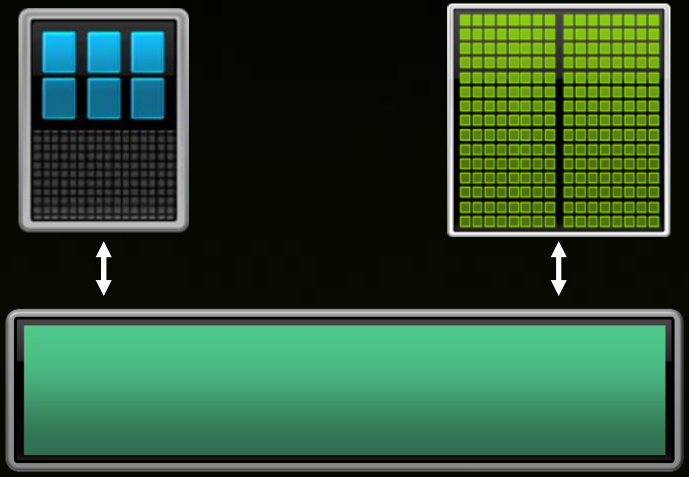
Developer View Today



System Memory

GPU Memory

Developer View With Unified Memory



Unified Memory

Simplified Memory Management in CUDA 6

CPU Code

```
void sortfile(FILE *fp, int N) {
  char *data;
  data = (char *)malloc(N);

  fread(data, 1, N, fp);

  qsort(data, N, 1, compare);


  use_data(data);

  free(data);
}
```

CUDA 6 Code with Unified Memory

```
void sortfile(FILE *fp, int N) {
  char *data;
  cudaMallocManaged(&data, N);

  fread(data, 1, N, fp);

  qsort<<<...>>>(data,N,1,compare);
  cudaDeviceSynchronize();

  use_data(data);

  cudaFree(data);
}
```

Roadmap eventually replaces cudaMallocManaged() with malloc()

Where are we going? [Technology]

- Hardware
  - Heterogenous CPUs -- x86, ARM, Power
  - NVLINK to ARM, Power for processor speed access to system memory
  - On package memory for Higher bandwidth, better density, more capacity
  - Unified Memory – easier to use
  - More parallelism!
- CUDA
  - More features in common languages like Java, Python
  - More libraries especially in machine learning, big data
  - C++17 proposed standards for parallel libraries (similar to Thrust)

Where are we going? [Finance]

- Traditional Markets
  - Real Time non-linear risk & margining
  - Larger/more complex baskets of underlyings
  - Higher dimensional models for PDEs
  - Non-gaussian/empirical models
  - Changes to the way we batch work

- New Markets
  - Model Risk – “multi-model” monitoring
  - Real time streaming CUSTOMER CENTRIC analytics
  - Geospatial models (Insurance and Fraud)
  - Generally Big Data & Deep Learning!

Recap – GPU Accelerated Compute in Finance

- Where did we come from?
  - Bleeding edge developers and IT pioneers delivering faster pricing & cheaper risk

- Where are we?
  - Packaged solutions and libraries plus improved productivity & performance tools in multiple languages combined with off-the-shelf IT solutions delivering faster & cheaper CVA, risk, and backtest

- Where are we going to?
  - GPUs will become even easier to own
  - New mathematical techniques, financial and customer models will grow to the available performance
  - Packaged solutions, libraries, and languages bring acceleration within reach for every firm
  - Customer centric analytics (“big data” coupled with machine learning)

Select web resources

- NVIDIA Computational Finance
  http://www.nvidia.com/object/computational_finance.html

- GTC Express Webinars
  http://www.gputechconf.com/resources/gtc-express-webinar-program

- GTC On Demand Presentations
  http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php

Selected web resources

- National University of Singapore Risk Management Institute (Oliver Chen)
  http://www.rmi.nus.edu.sg/

- Dalhousie University Risk Analytics Lab (Andrew Rau-Chaplin)
  http://www.risk-analytics-lab.ca/

- Oxford University (Mike Giles)
  http://www.maths.ox.ac.uk/people/profiles/mike.giles

- NUS Risk Management Institute
  http://www.rmi.nus.edu.sg/

- University of Melbourne / QuantLib & Kooderive (Mark Joshi)
  http://www.markjoshi.com/ & http://sourceforge.net/projects/kooderive/

Selected web resources

- Napier 400 http://www.royalsoced.org.uk/cms/files/events/programmes/2013-14/Draft%20napier%20programme.pdf

- University of Chicago “Recent Developments in Parallel Computing in Finance” https://stevanovichcenter.uchicago.edu/page/recent-developments-parallel-computing-finance

- WHPCF13 http://portalparts.acm.org/2540000/2535557/fm/frontmatter.pdf?ip=62.216.237.3&CFID=501111212&CFTOKEN=55864985

- Call for papers WHPCF14 http://ewh.ieee.org/conf/whpcf/

- Global Derivatives http://www.icbi-derivatives.com/