Market Landscape For AI Hardware

Karl Freund

karl@moorinsightsstrategy.com

Sr. Analyst, AI and HPC

Moor Insights & Strategy

Follow my blogs covering Machine Learning Hardware on Forbes: http://www.forbes.com/sites/moorinsights or https://muckrack.com/karl-freund/articles

www.MoorInsightsStrategy.com

GOALS OF THIS SESSION

1. Lay the foundation for the next two days
2. Review AI Strategy from the major vendors
3. Introduce you to a few new companies

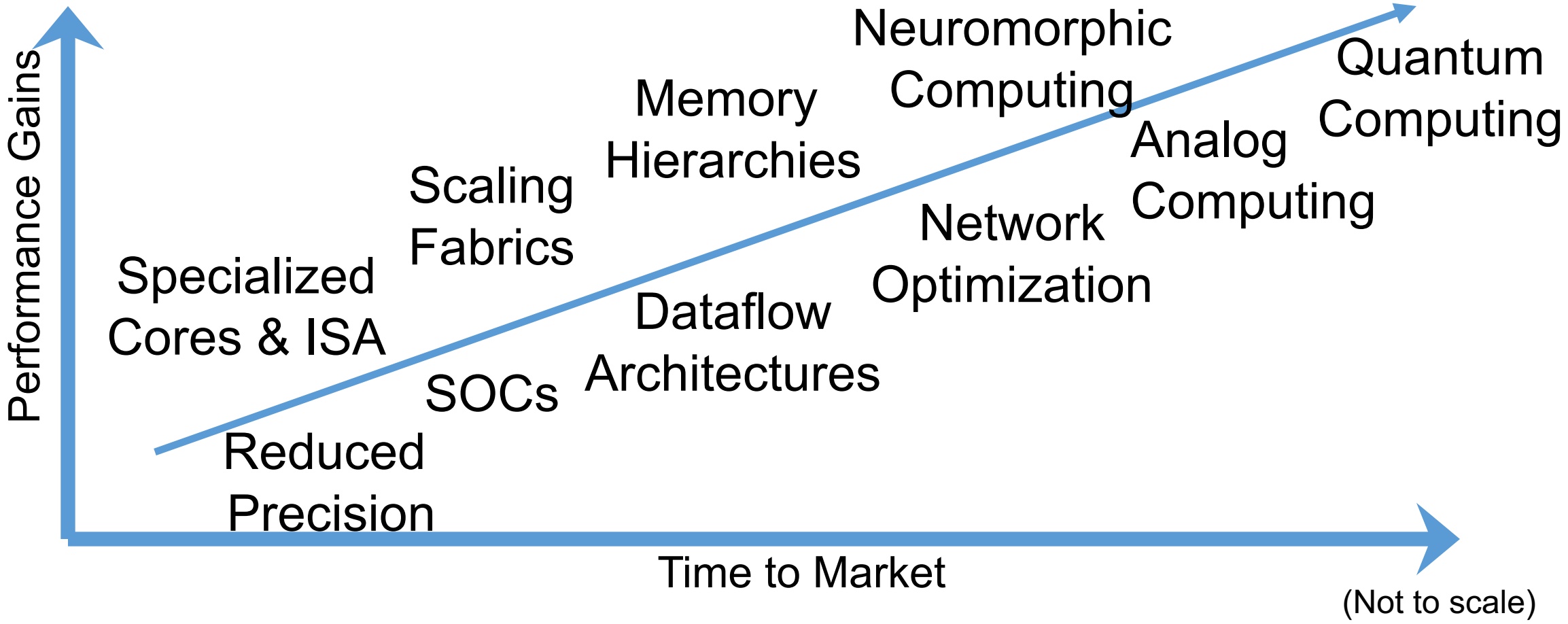
The Two Worlds of AI Processing

1. **Training a Network**
   - Runs take days or even weeks
   - Trilions of billions of ops
   - Massive data sets
2. **Inference Processing**
   - Runs take milliseconds
   - Huge range of requirements
   - Will become the larger market

HARDWARE CHOICES FOR AI (DEEP LEARNING)

- **CPUs**: Good for inference and model prototyping
  - Use the CPU's you already have
- **FPGA's**: Very good for Inference and other tasks (See Microsoft)
  - Difficult to master.
- **GPU's**: Very flexible accelerators.
  - Great for training and demanding inference workloads (Conversational AI)
  - Programmable for new models, new networks
  - Massive ecosystem already in place
- **ASIC's:** Great potential in AI, however …
  - Requires significant R&D and time; need high volumes to break-even
  - Entails long development cycles in a fast-moving ecosystem
  - Each ASIC will need an ecosystem of software, models, researchers
  - Not a fungible asset in the cloud

IMPROVING AI PROCESSING



THE COMPETITIVE LANDSCAPE

"*A CAMBRIAN EXPLOSION IN CHIPS*"

HTTPS://WWW.FORBES.COM/SITES/MOORINSIGHTS/2019/01/23/2019-A-CAMBRIAN-EXPLOSION-IN-DEEP-LEARNING-PART-1/#3CE8172F4DC1

NVIDIA:
LEADER IN TRAINING, ALSO TARGETING INFERENCE

- GPU's benefit from programmability
  - –Mature SW stack and development platform
- Volta delivers industry-leading 125 TOPS today
- T4 GPU is gaining inference traction (Google & AWS)
- Xavier SOC : flexible platform for edge AI
  - –"DLA" is open-source ASIC for CNNs
- All clouds, OEM's, Universities, SW Startups



INTEL HAS A LOT OF AI FIREPOWER

- New Xeons w/ 11x improvement in AI performance
- Nervana NNPs coming soon (training & inference)
- Acquired Mobileye for AVs and Movidius for edge vision
- Intel's future Exascale Engine (**X^{e}**) could be game changer
  - Very few details have been disclosed.

https://www.forbes.com/sites/moorinsights/2019/03/19/intel-and-cray-reaffirm-first-usa-exascale-supercomputer/#6bced6fc29df

QUALCOMM
(YES! QUALCOMM!)

- Strong legacy of performance, power efficiency, volume and AI technology
  - –Snapdragon now includes 4th generation AI Engine
- Qualcomm has announced their intentions to extend their presence in AI into the Data Center (Cloud AI100)
  - –"Distributed Intelligence" Vision
- Watch this space

https://www.forbes.com/sites/moorinsights/2019/04/10/qualcomm-dives-into-the-deep-end-of-the-data-center/#710d9163367f

FPGA's: Flexible, Programmable Inference

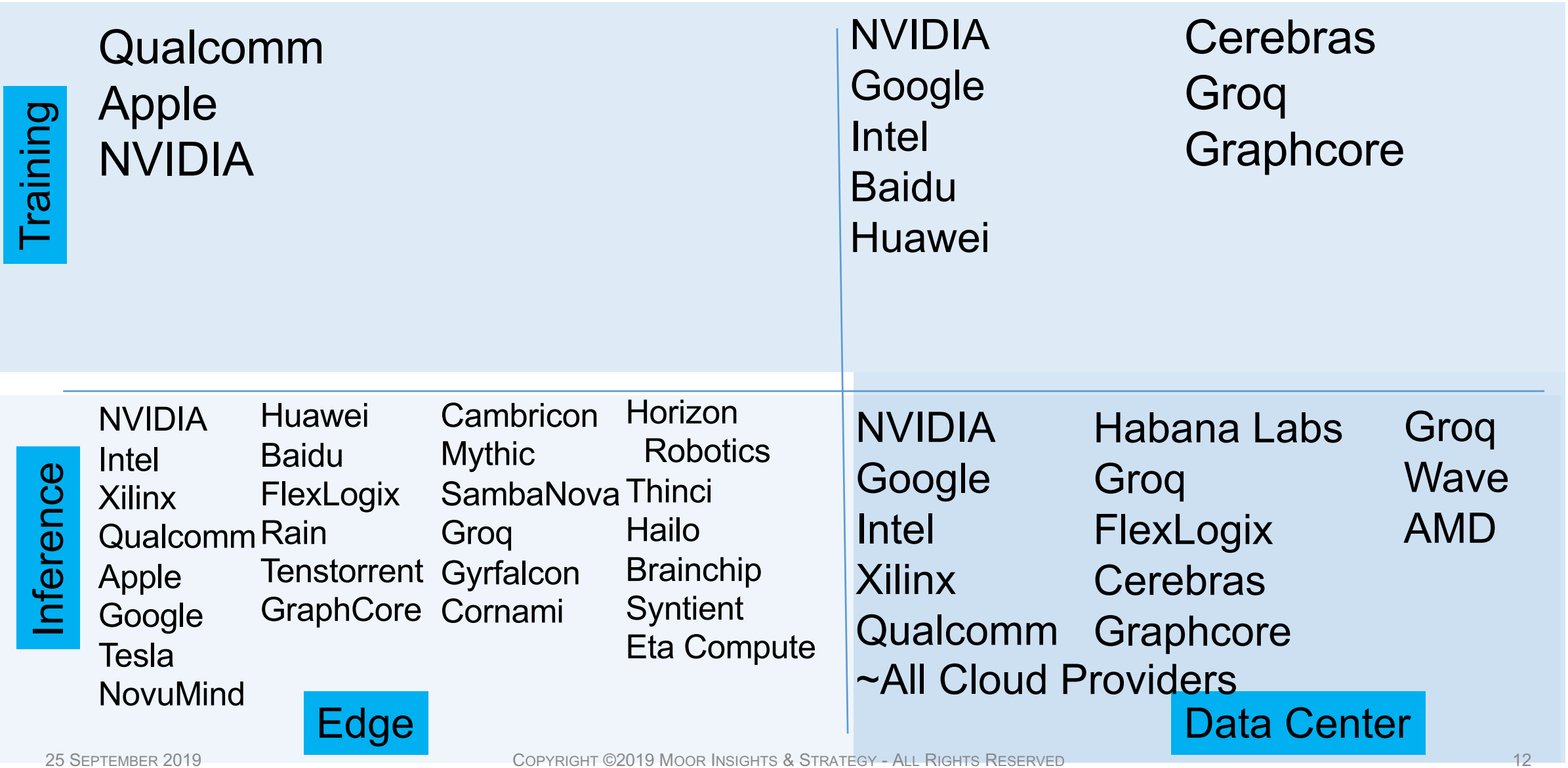
- HW programmable accelerators from Intel and Xilinx
- Microsoft champions FPGAs in datacenter
- Xilinx and AWS offer pre-built solutions
- AWS offers Xilinx as a Service
- FlexLogix is developing an Inference-specific FPGA (2020)

https://www.forbes.com/sites/moorinsights/2019/08/28/xilinx-reveals-more-versal-details/#21b1147e330d

AI Processing in Public Clouds

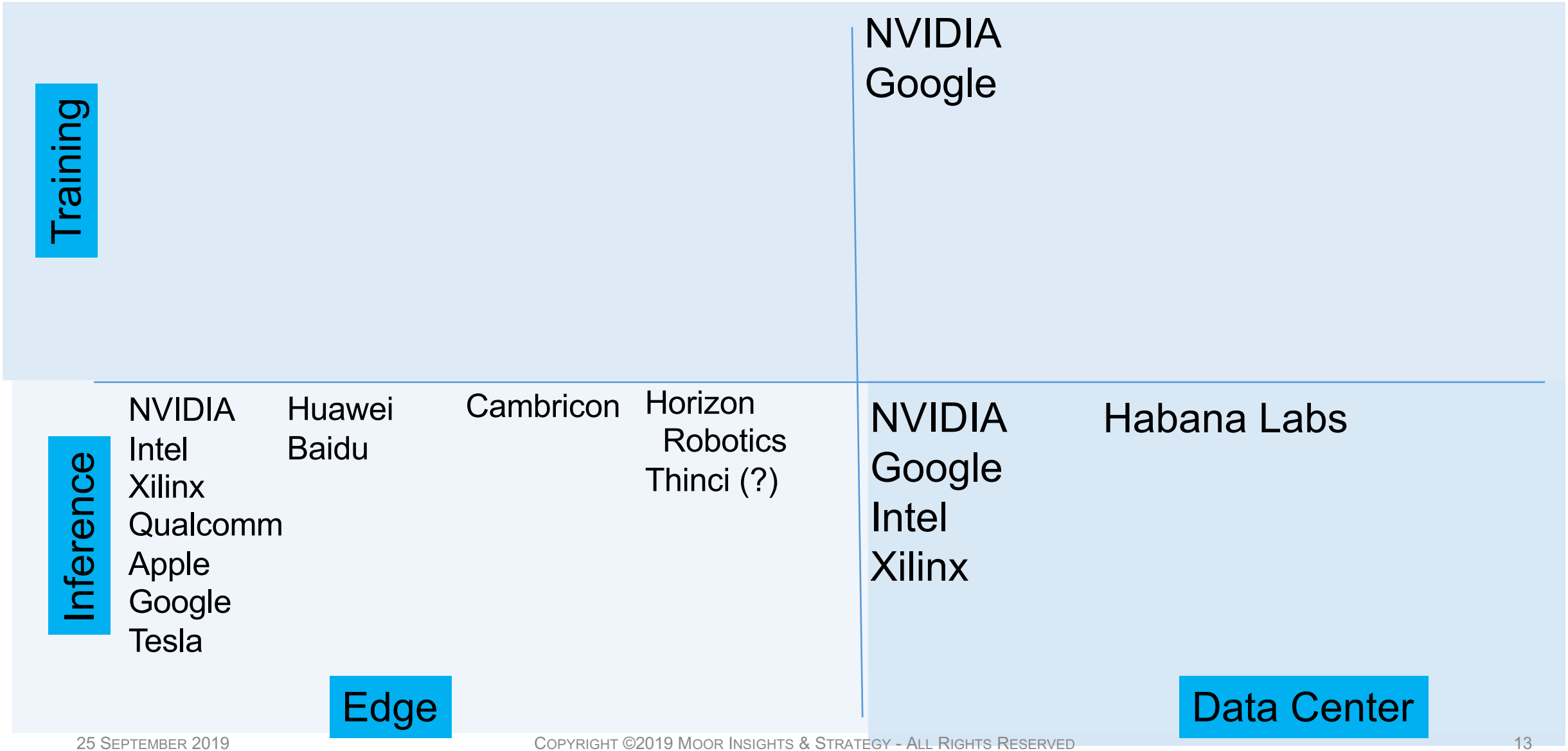
- AWS has the large public cloud AI portfolio
  - Offers extensive suite of AI development tools & networks, many to support Alexa
  - Is building their own Inference chip ("Inferentia")
  - Offers an app marketplace and F1 instances for Xilinx FPGAs
- Microsoft has invested heavily in GPUs and FPGA's
  - Extensive API library with pre-trained neural networks
  - FPGA's currently for internal use across MSFT infrastructure.
- Google has the TPU and a huge AI team
  - Tens of thousands engineers and scientists
- Alibaba and Baidu intend to build out indigenous Industry
  - Baidu has announced their own chips, Alibaba has founded a chip company

Company Target Markets (partial List)



CHIPS IN PRODUCTION*

* Does not Imply Adoption



FINALLY, A FEW THOUGHTS…

1. TOPS = **T**remendously **O**verused **P**erformance **S**tat
   - Network specific benchmarks are better
   - Mlperf may become the solution, all vendors should publish!
2. Data Center AI demands programmability
3. Never, ever underestimate NVIDIA (or INTEL)
   - Nobody saw TensorCores coming.

STAY TUNED FOR WHAT'S NEXT! @karlfreund

THANK YOU!

KARL FREUND, SR. ANALYST, AI AND HPC
MOOR INSIGHTS & STRATEGY
KARL@MOORINSIGHTSSTRATEGY.COM
@KARLFREUND

WWW.MOORINSIGHTSSTRATEGY.COM