SYSTEM DESIGN FOR LEARNING MACHINES

NICHOLAS MALAYA

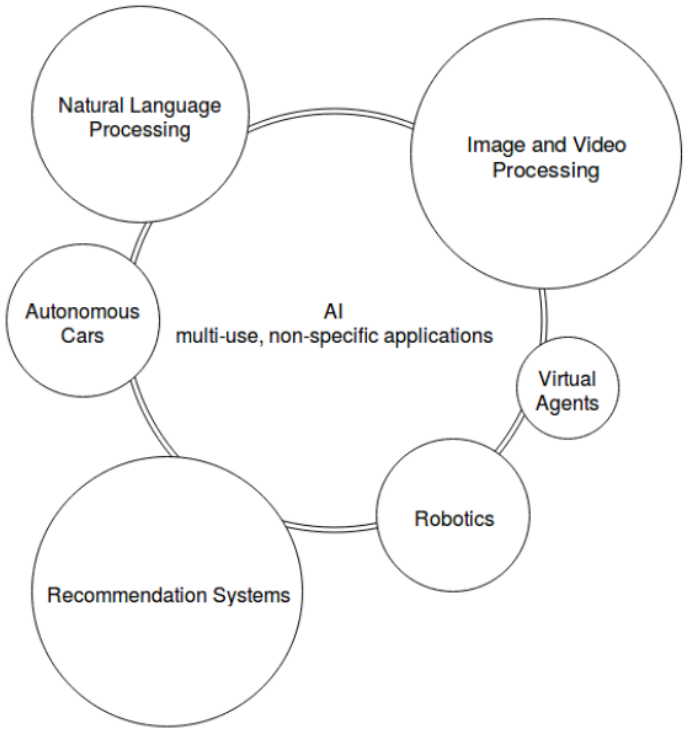
AMD RESEARCH

11/28/2018

GROWTH

- McKinsey Global Institute Report:
  - 2.2 Exabytes of data every day

- $18- $27 Billion Internal
- $1.5B Baidu R+D
- $1B – Toyota autonomous cars
- $8- $12 Billion External
- 2-3% of all VC funding
- **$59.8 Billion market by 2025**



AMD: ADVANCED MICRO DEVICES

- CPUs [Ryzen]
  - One of two x86 suppliers
- GPUs [Radeon]
- Gaming Consoles:
  - Xbox One X, PS4 Pro, WiiU
- APUs, Servers [EPYC], Supercomputers, etc.

Blue Waters Supercomputer

Range of hardware in many domains

WHY MACHINE LEARNING?

"KILLER APPLICATION" OF GPUS

Simple answer: interested in all computation

Machine Learning is particularly interesting

- Potential to impact nearly all software/industries
  - NLP, autonomous cars, etc.
  - "Software is eating the world, AI is eating software"
- Compute intensive (good)
- Amenable to acceleration / specialized hardware

ESSENTIAL CHALLENGE

- **Today the job of training machine learning models is limited by compute,** if we had faster processors we'd run bigger models...in practice we train on a reasonable subset of data that can finish in a matter of months. **We could use improvements of several orders of magnitude - 100x or greater.**

-Greg Diamos, Senior Researcher, SVAIL, Baidu
September 27, 2016

HOW LONG DOES THIS TAKE?

TRAINING

ResNet-50:

- 53 hidden CONV layers
- Top-5 error: **5.3%** (on **image-net**)
- Human: **5.1%**

Image-net:

- "Standard benchmark" -- 1.2 million images (training) --14,197,122 annotated images
  - Largest such dataset in the world

3 days to train ResNet-18 for 30 epochs with four GPUs

- **Tesla K80 (224 GB)**
- **Top-5 accuracy: 91.97%**

ESSENTIAL CHALLENGE

- **We could use improvements of several orders of magnitude**

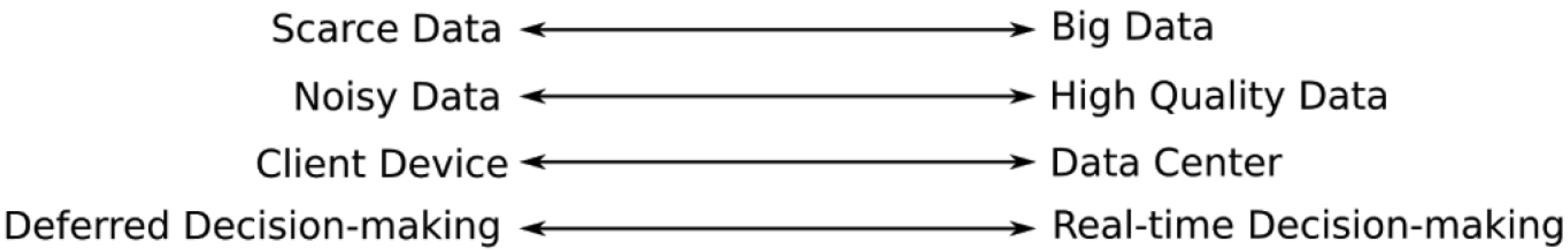
- Where could these improvements come from?
  -
  -
  -
  -

ESSENTIAL CHALLENGE

- **We could use improvements of several orders of magnitude**

- Where could these improvements come from?
  - Reduced compute (quantization, sparsity)
  - Algorithms (GANs, VAE, etc.)
  - Parallelism (Data and Model)
  - Specialized Hardware (Dennard Scaling is Dead)

RANGE OF APPLICABILITY



- DL is not a single solution

REDUCED PRECISION

REDUCED PRECISION MOTIVATION

- Why do we care?
  - Desire to reduce the amount of computation

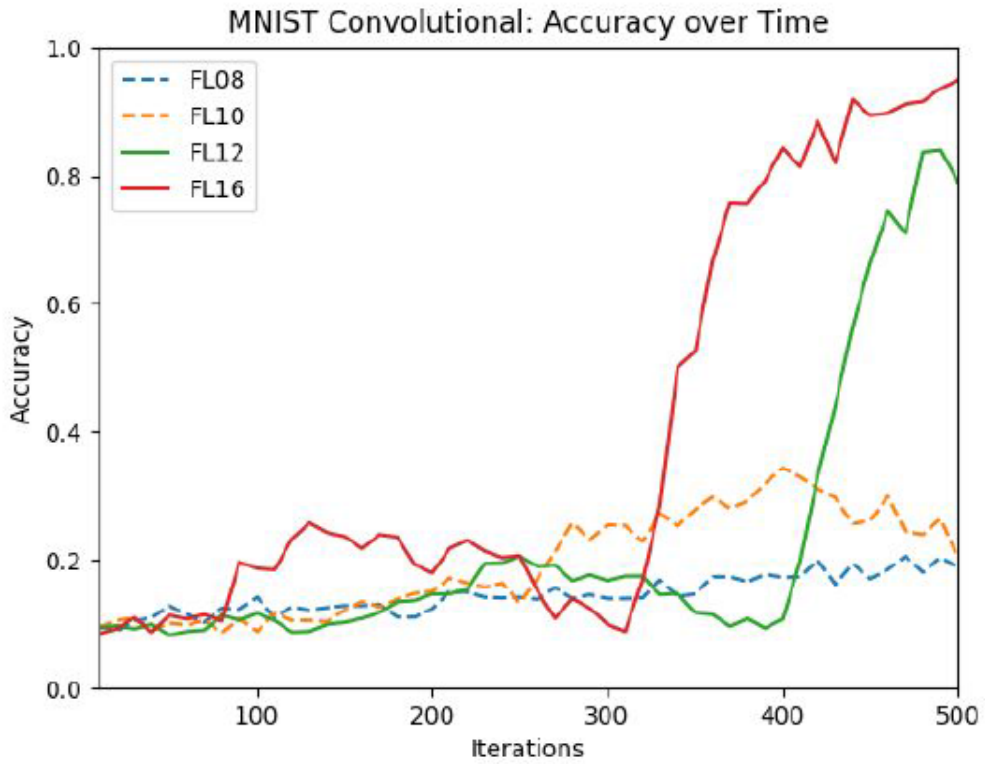
- Consider the AMD Vega20 GPU
  - **7.4 TFLOPs double precision (FP64) compute**
  - **14.8 TFLOPs single precision (FP32)**
  - **29.5 TFLOPs half precisions (FP16)**

  - Similar numbers on other hardware devices

HOW LOW CAN WE GO?

- Train with 16-bit and stochastic rounding (Gupta, et al. 2015)
- Train with low-precision multiplications (Courbariaux, et al. 2014)
- Train with binary weights (Courbariaux, et al. 2015)
- One-bit gradient for parallelization of SGD (Seide, et al. 2014)

REDUCED PRECISION MOTIVATION



REDUCED PRECISION MOTIVATION

- No previous work capable of predicting sensitivity to precision
- No estimate of precision tolerance established
- No predictions of how bit size affects neural network

- More theory is needed in the field.

DERIVING A STABILITY BOUND ON FWD PROP

Join work with Naman Maheshwari, Scott Moe, Sudhanva Guruthimuthi

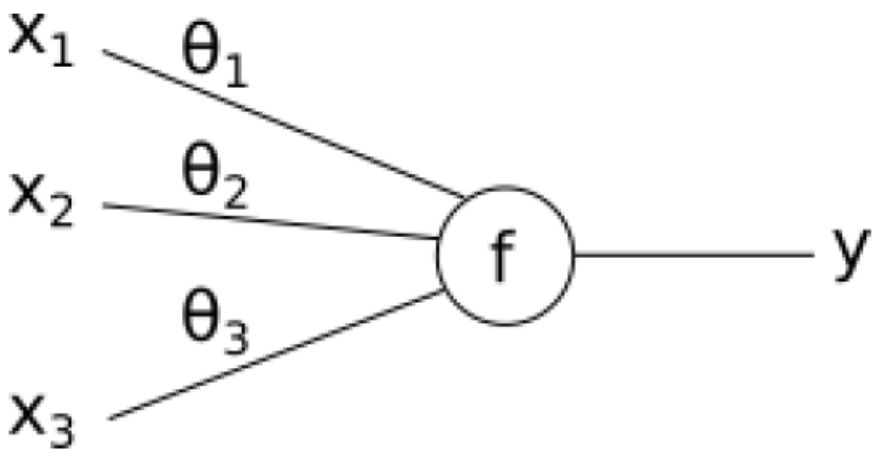
- Single Neuron:

$y = f(\theta_i x_i)$

$y = f(\theta_i x_i) < (\theta_i x_i)$

$||y||\ \ ||\theta_i||\ ||x_i||$

$$\frac{\|f(\theta\, x)\|}{\|x\|} \leq \|\theta\|$$



GENERALIZED TO AN N-LAYER NETWORK

$$\frac{\|f(\theta\, x)\|}{\|x\|} \leq \|\theta\|$$

Can be generalized to:

$$\frac{\|\delta_y\|}{\|\delta_x\|} \leq \sum_{i=1}^{n} \left(\prod_{j=0}^{i-1} \theta_{n-j} \right)$$

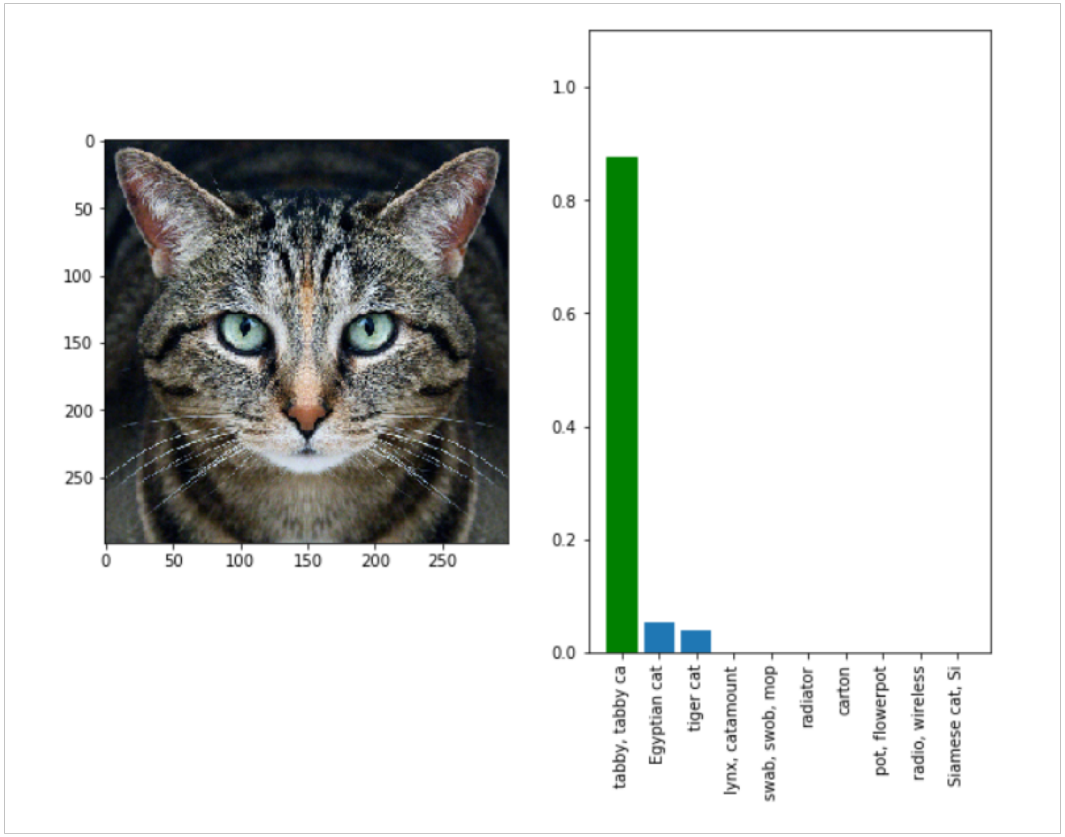
REDUCED PRECISION MOTIVATION

- Analogous to the “condition number” in numerical linear algebra
- Criterion for stability:

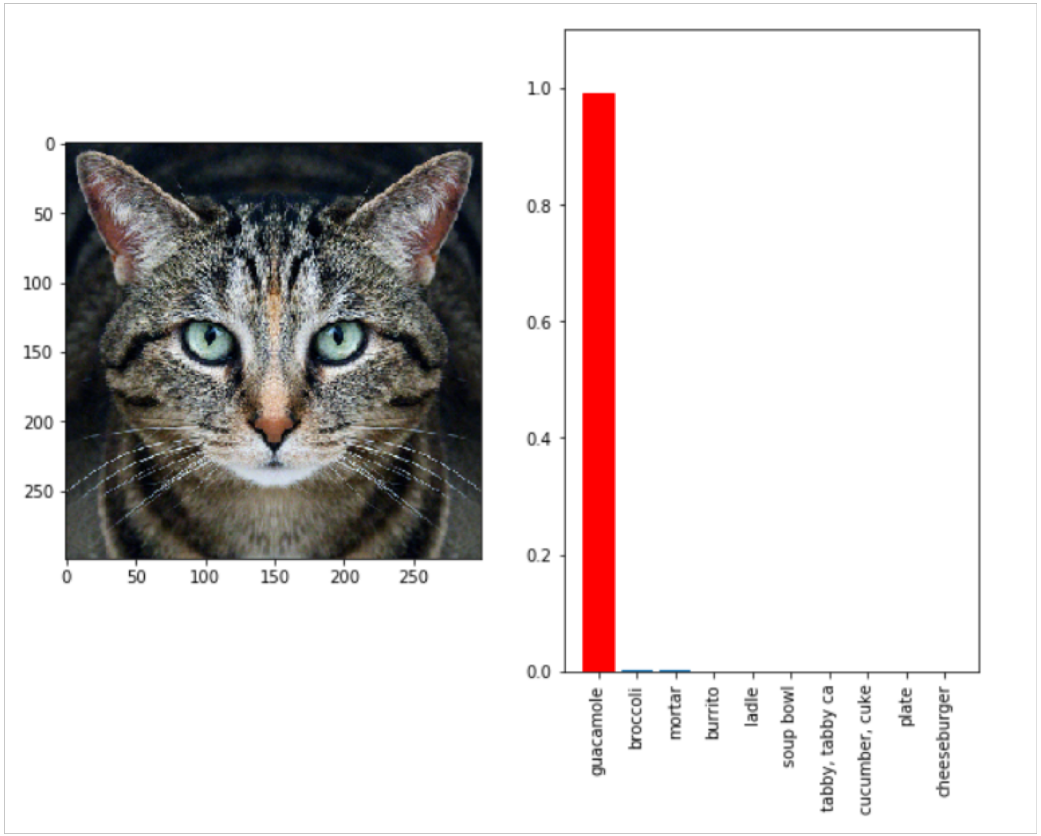
$$\kappa_N^F \epsilon < O(1)$$

- Remaining question: how tight is this bound?
  - Not all perturbations are equal.
  - Adversarial attacks

TABBY CAT?



NOPE; GUACAMOLE



ADVERSARIAL LEARNING

- What happened?
  - 1) Misclassification
  - 2) Small perturbation to data

- Why is this bad?
  - Indicates models are not robust
  - Security threat:
    - Fool an autonomous vehicle by modifying traffic signs
    - Turtle, or gun?

      - **Can adversarial learning point to more robust training?**

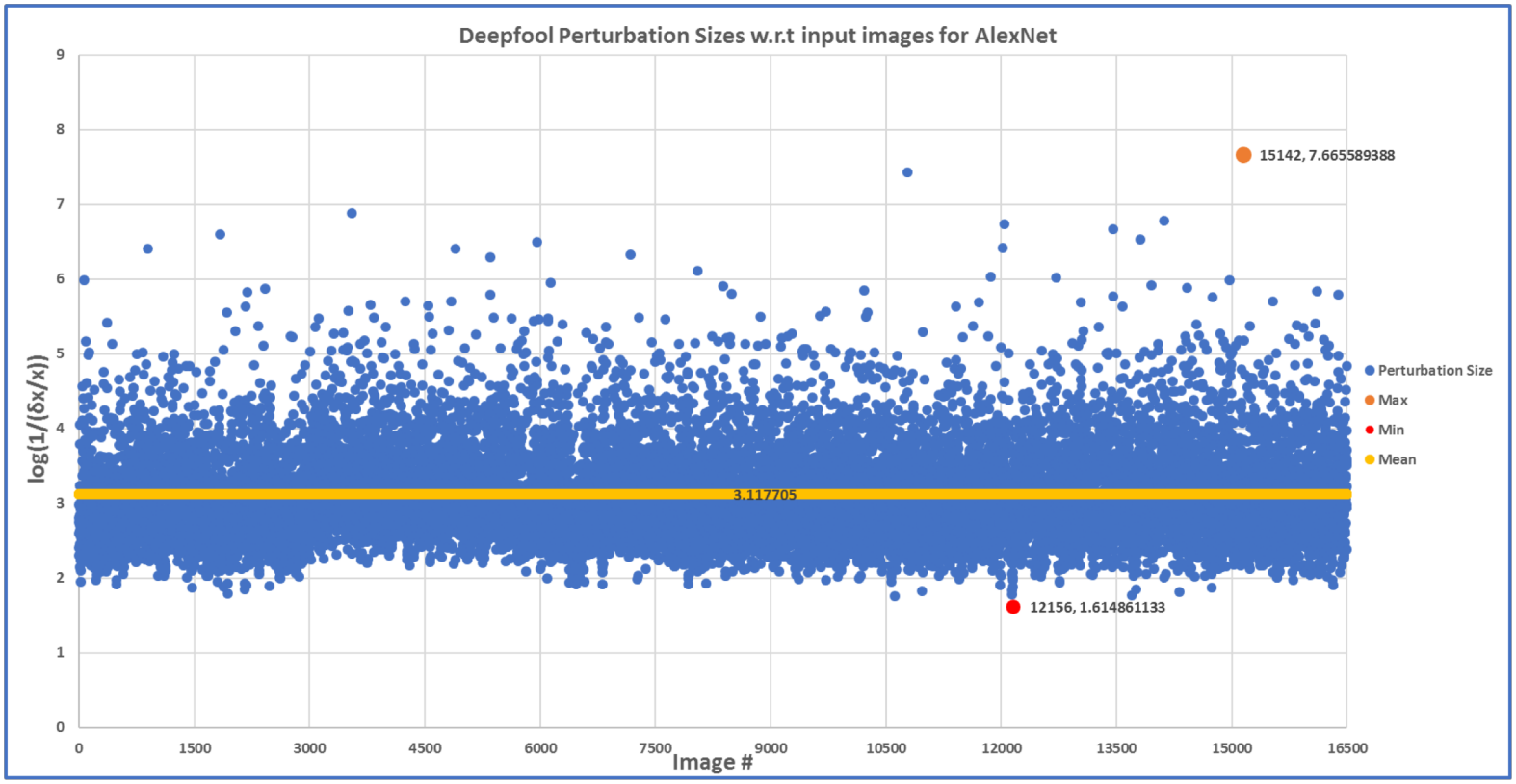
ADVERSARIAL ATTACKS AS PERTURBATIONS

•Various methods to generate adversarial attacks: deepfool, FGS
•FGS: linearize cost function around the current value of θ, obtaining an optimal max-norm constrained perturbation,

$$\eta = \epsilon \, sign\big(\nabla_x J(\theta, x, y) \big)$$

•Is this optimal? What do you know about SGD?

RESULTS OF THE ESTIMATOR



Predicted
10e9

AMD

IMPLICATIONS ON SYSTEM DESIGN

Hardware:

- What precisions should be supported?
- What hardware technologies are (im)possible based on this analysis

Software:

- Precision support
- Estimate max evaluation error
  - e.g., certification / validation / uncertainty quantification

Architectures and Algorithms:

- Which DNN architectures are resilient to reduced precision?
- What future architectures are more/less likely?

FUTURE WORK

Characterizing precision tolerances across range of ML workloads / applications

- E.g., CANDLE applications (HPC + ML)
- CNNS (resnet, VGG, inception),
- RNNs (seq2seq),
- Reinforcement Learning (A3C),
- GANs/VAEs, etc.
- Also pruned networks (squeezenet, tiny_yolo)

Similar results available for backprop

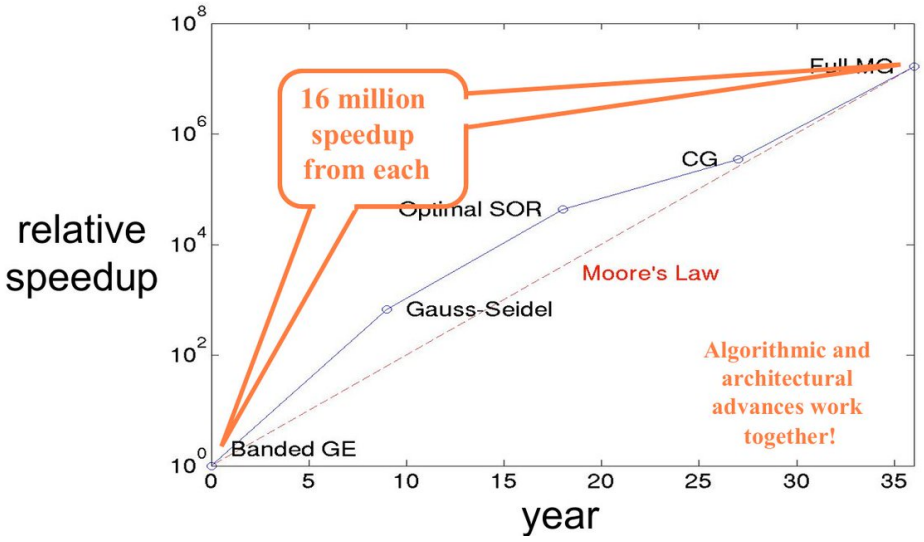
(I HAVEN'T EVEN ADDRESSED SPARSITY)

ALGORITHMS / GANS

ADVANCED ALGORITHMS MOTIVATION

Algorithms and Moore's Law

- This advance took place over a span of about 36 years, or 24 doubling times for Moore's Law
- $2^{24} \approx 16$ million ⇒ the same as the factor from algorithms alone!

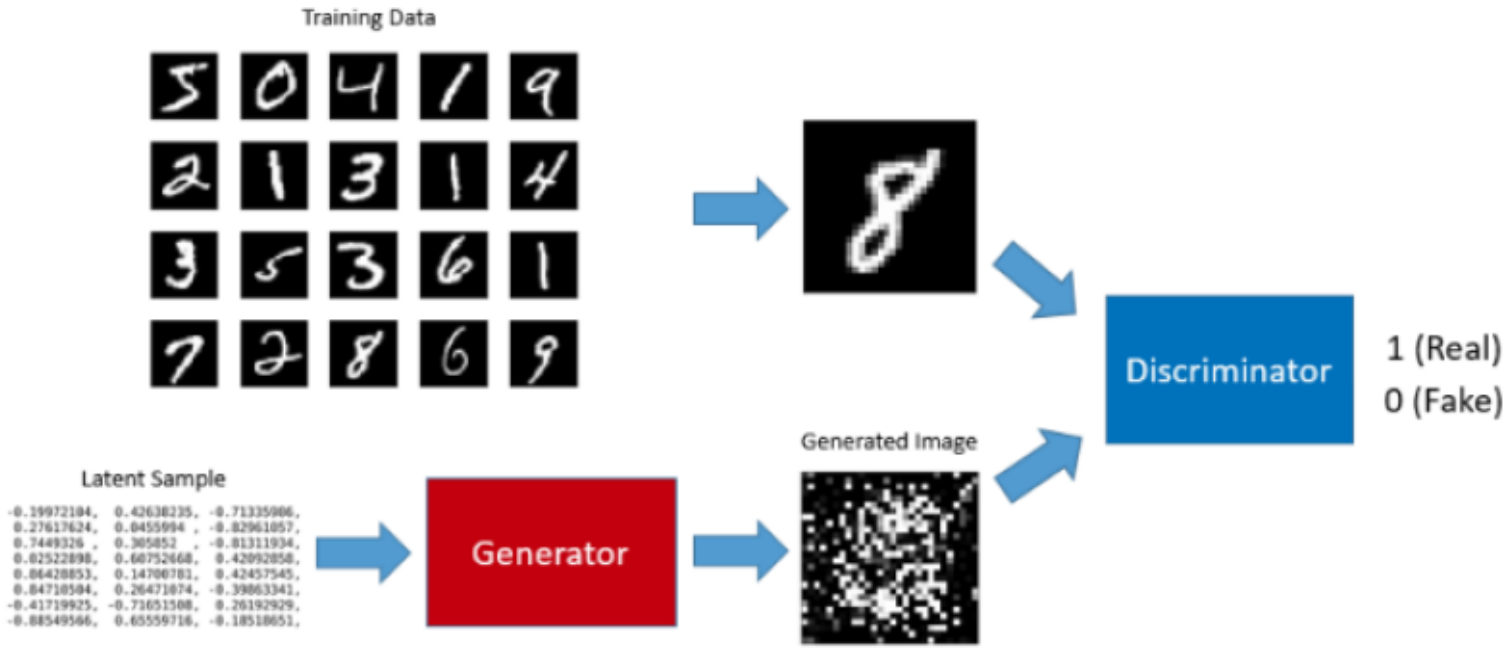


- Slide from David Keyes' "Algorithmic Adaptations to Extreme Scale Computing"
- "Algorithmic and architectural advances work together"
- 36 year period ~ same factor
- Will continue to Exascale
- "Software and architectural advances work together"

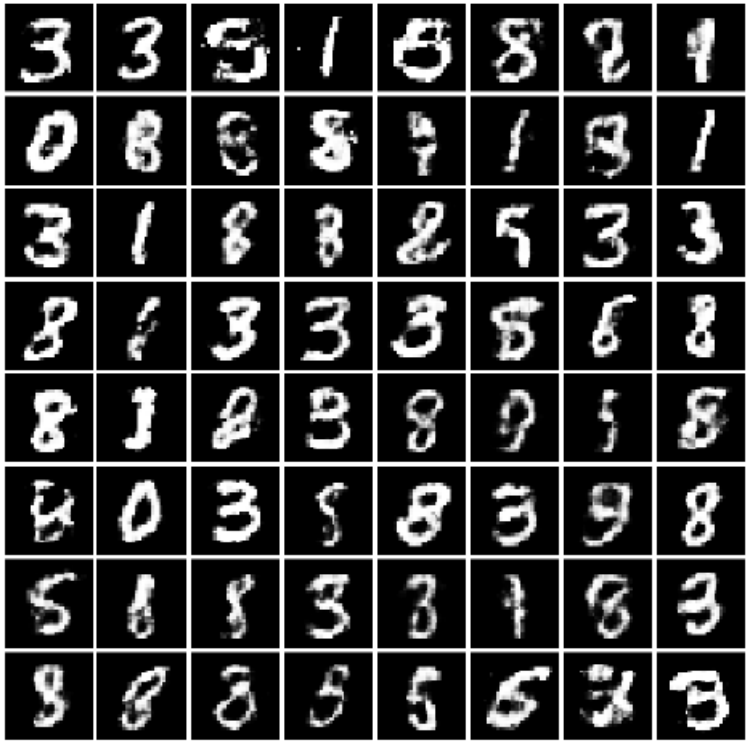
AMD

GANS / VAES

LEVERAGE EXISTING MACHINERY



TEACHING AN AI TO WRITE



This GAN learns to ‘write’ the digits 0-9

GAN learns many digits pretty well,
it never really learned 7!
Not many good examples of 4s either.

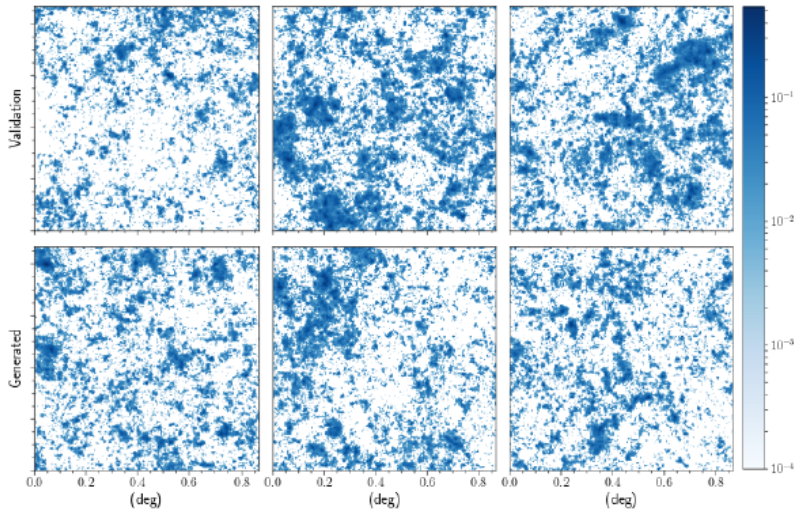
So this is an example of an AI that
Superficially learns to write, but is
Missing key flavors from what even a
Little child would know.

FAKE CELEBRITIES

Legal: Steal someone's likeness
Security implications
Fake news!

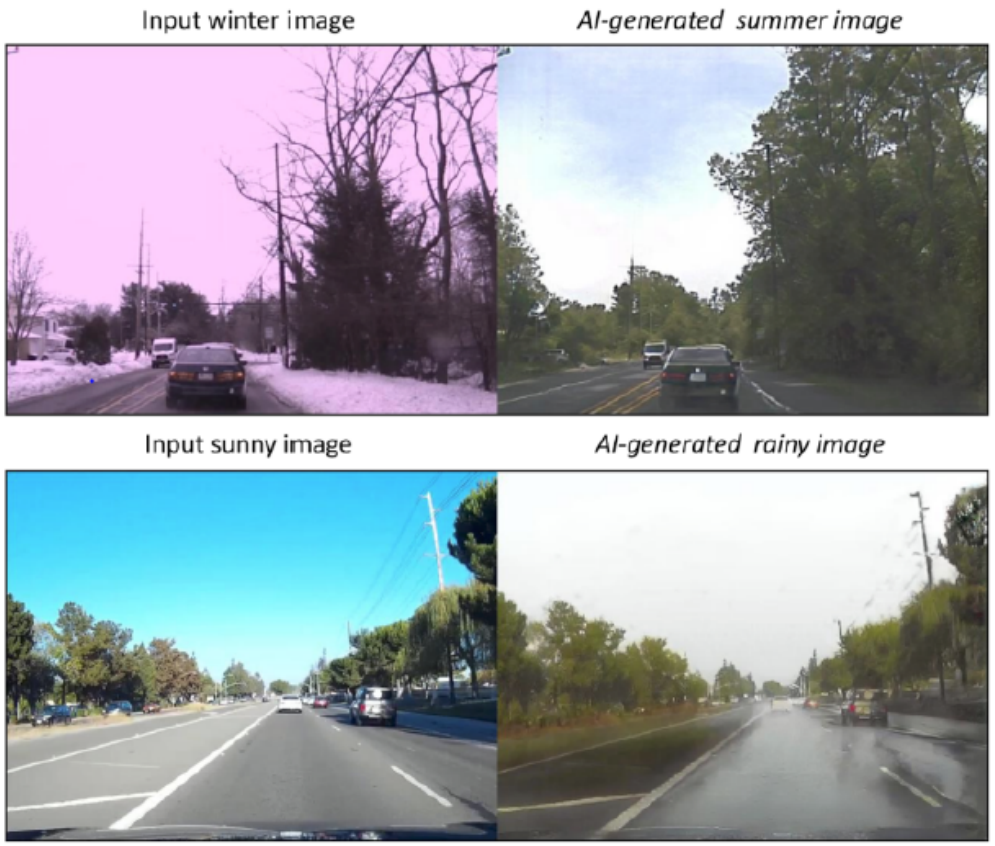
"Progressive Growing of GANs for Improved Quality, Stability, and Variation" (ICLR 2018)

GAN USES



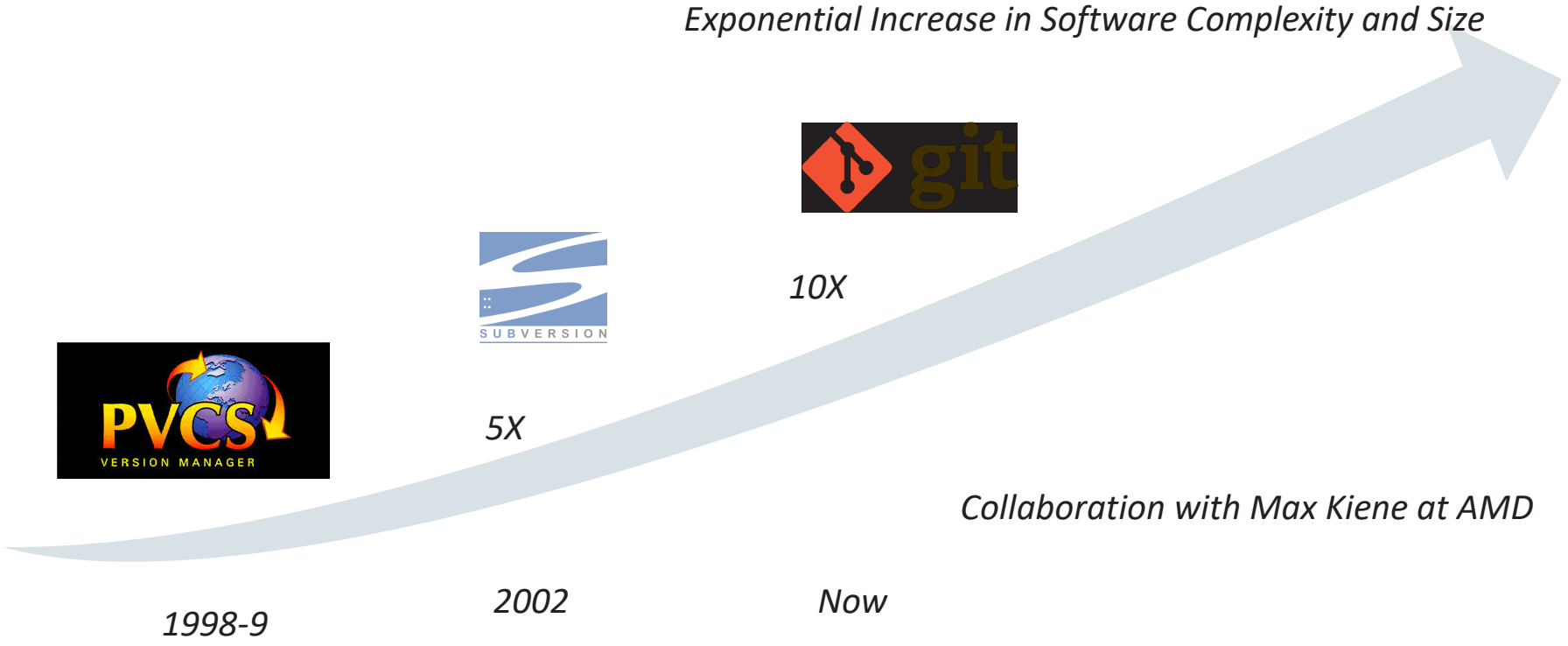
- NERSC CsomoGAN: Cosmology Mass Maps
- Accurately represents energy spectrum: Universal Approximators
- using DCGANs (the architecture shown previously)

GAN USES



- Generate data for different scenarios

GAN USES: DEFECT PREDICTION



AMD

DEFECT PREDICTION

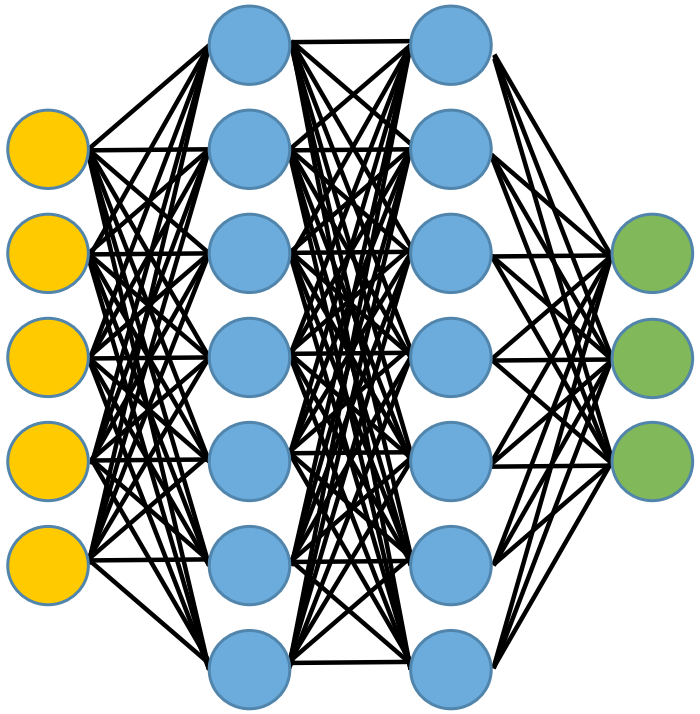
- Classifier for incorrect check-ins
- Predict where our tests are insufficient
- GANs: automatic test generation?
- Many companies are not sitting on huge troughs of data

SCALE-UP / SCALE-OUT

SCALE-UP/SCALE-OUT

- Scale-up:
  - "Fat Nodes"
  - Density of compute (but also power demands)
- Scale-out:
  - More traditional "HPC"
  - Requires network communication

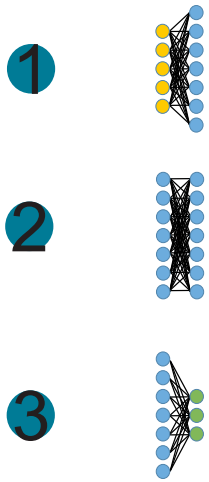
PARALLELISM IN DEEP LEARNING



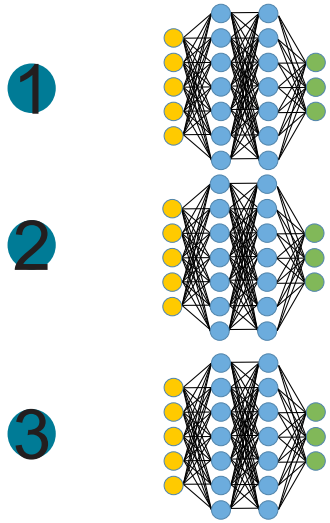
Joint work with Abhinav Vishnu

PARALLELISM TYPES

WHICH TYPES OF DL PARALLELISM MAKES SENSE? AND WHEN?



Model Parallelism



Data Parallelism

PARALLELISM IN DEEP LEARNING

- Why does communication happen?

PARALLELISM IN DEEP LEARNING

- Why does communication happen?
  - In inference? *Nowhere, or only to distribute data, model, etc.*
  - What about training?

PARALLELISM IN DEEP LEARNING

- Why does communication happen?

$$\theta' = \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$$

  - θ – Weights; η – Learning rate; J – Cost function

PARALLELISM IN DEEP LEARNING

- Why does communication happen?

$$\theta' = \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$$

  - θ – Weights; η – Learning rate; J – Cost function

$$\frac{\partial J(\theta)}{\partial \theta} \approx \frac{1}{N} \sum_{i} \frac{\partial J(\theta_i)}{\partial \theta_i}$$

PARALLELISM IN DEEP LEARNING

$$\frac{\partial J(\theta)}{\partial \theta} \approx \frac{1}{N} \sum_i \frac{\partial J(\theta_i)}{\partial \theta_i}$$

- What is an averaging operation?

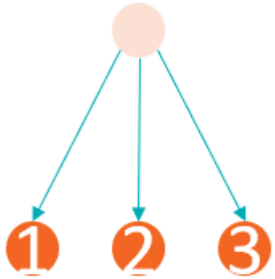
$$\sum_i a_i = a_1 + a_2 + a_3 + \ ... + a_N$$

Just an *all-reduce!*

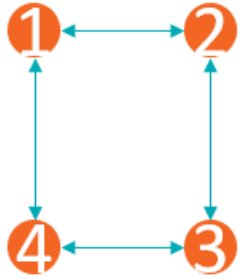
PARALLELISM IN DEEP LEARNING

- Where communication happens:
  - Parameter server (*Dean et al.)*
  - Ring-based all-reduce

- What are the pro/cons?



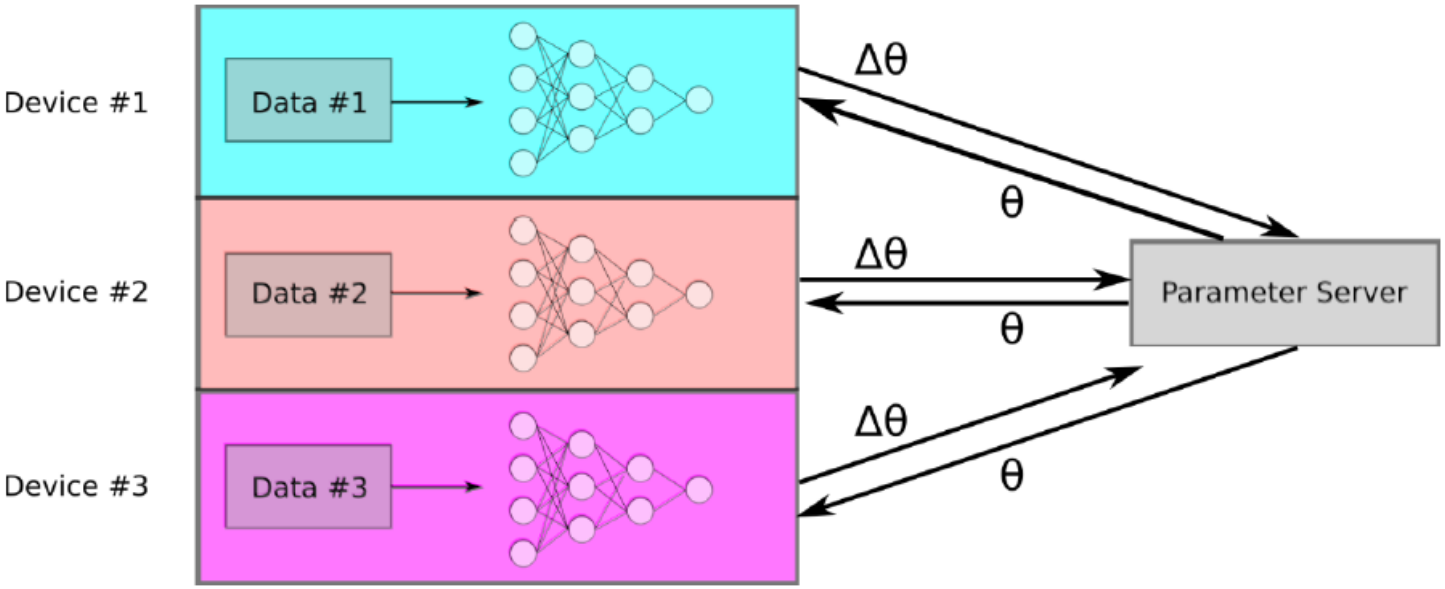
Parameter
Server based



Non-Parameter
Server based

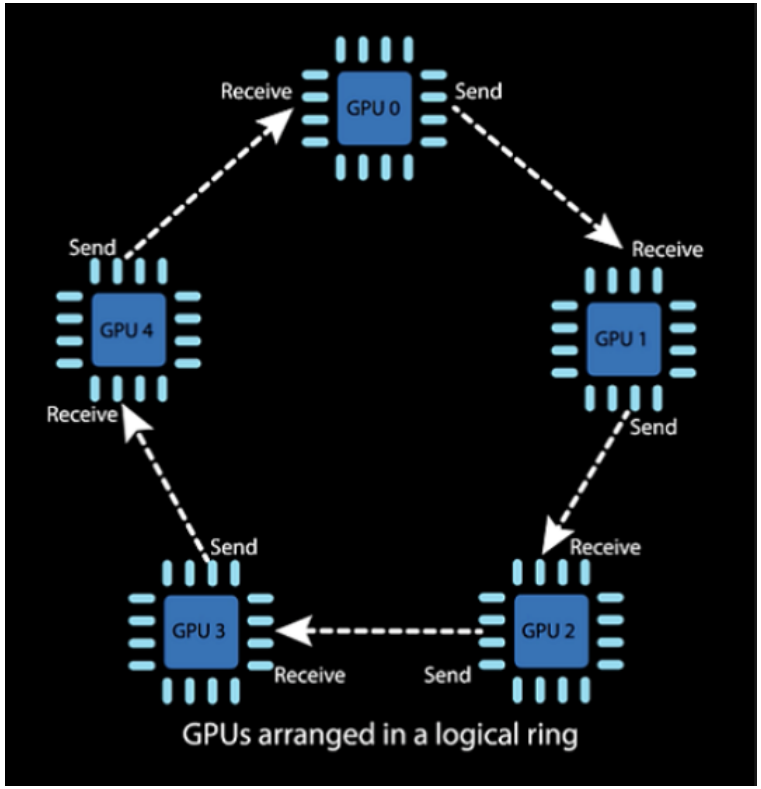
Compute node/device

PARAMETER SERVER (DEAN ET AL.)



- Push ($\Delta\theta$) and pull (θ) after every iteration
- Network Contention every iteration

RING



- Push ($\Delta\theta$) and pull ($\Delta\theta$) after every iteration

PARALLELISM IN DEEP LEARNING

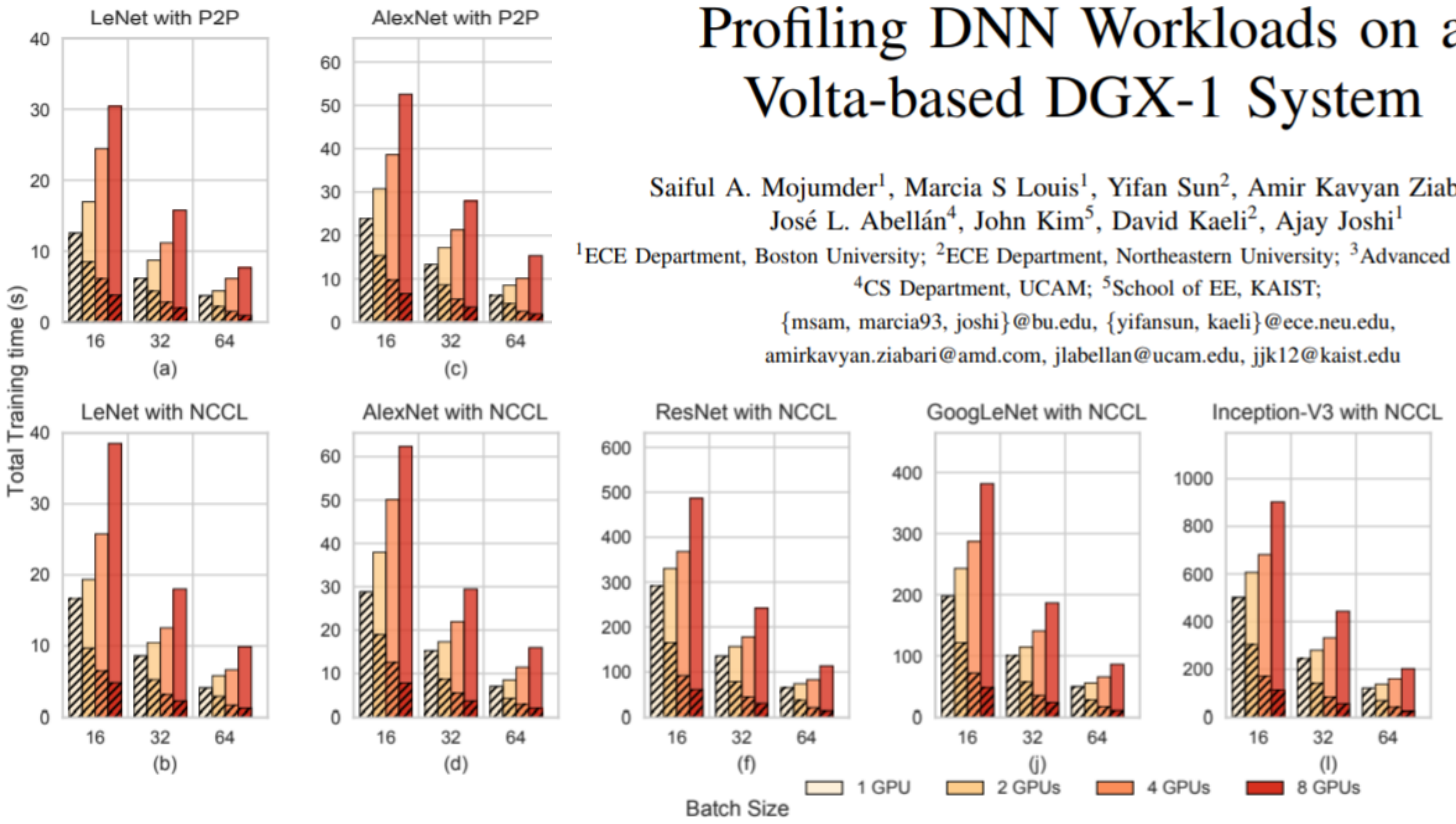
- How does communication happen?
  - PCIe: **hipdevicecopy(), cudaMemcpy()**
  - Distributed: **MPI_Allreduce**
  - On Node: **ncclAllReduce() / rcclAllReduce()**
  - In your favorite framework: gRPC (TF), Gloo (Caffe2)

FAST INTERCONNECT EXAMPLES

- **Seq2seq**
  - 7 layer deep NLP (Fully connected)
  - **2.7x speed-up over (4x) infiniband**

- **Mixture of Experts (MoE):**
  - From Google.
  - MoE layers each consist of 128 experts, each is feed-forward DNN.
  - Each expert specializes in a different domain of knowledge,
  - Experts distributed to different GPUs,
  - significant all-to-all traffic due to communications
  - Training dataset is“1 billion word benchmark”
  - Batch size of 8,192 per GPU.
  - **2x improvement over (4x) infiniband**

COLD WATER



Profiling DNN Workloads on a Volta-based DGX-1 System

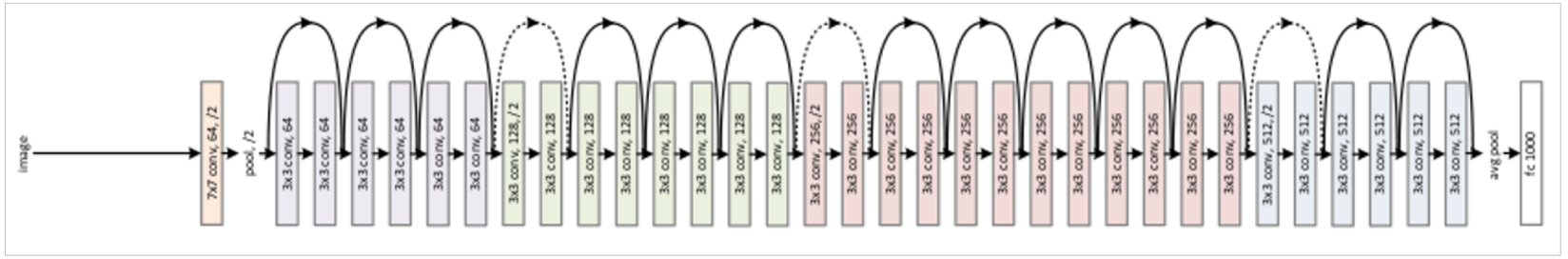
Saiful A. Mojumder[1], Marcia S Louis[1], Yifan Sun[2], Amir Kavyan Ziabari[3*],
José L. Abellán[4], John Kim[5], David Kaeli[2], Ajay Joshi[1]

[1]ECE Department, Boston University; [2]ECE Department, Northeastern University; [3]Advanced Micro Devices;
[4]CS Department, UCAM; [5]School of EE, KAIST;
{msam, marcia93, joshi}@bu.edu, {yifansun, kaeli}@ece.neu.edu,
amirkavyan.ziabari@amd.com, jlabellan@ucam.edu, jjk12@kaist.edu

AMD

USECASE: RESNET-50

- Image Classification
- Many 3x3 convolutions (50 layers deep)
- “Standard Benchmark” of Deep learning
- 4x GPUs working in data parallelism
- ResNet50 gradients are 120MB in FP32

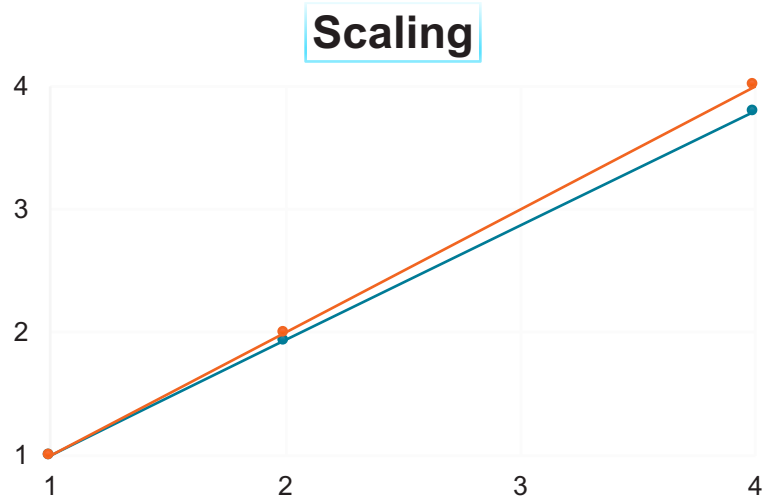


AMD

FAST INTERCONNECTS

- "Infinity-Fabric"
  - PCIe Gen3/Gen4:
    - Unidirectional bandwidth: 16GB/s (x16), 12 GB/s (effective)
  - Infinity-Fabric:
    - Unidirectional bandwidth: 50 GB/s, 40 GB/s (effective)
  - Infinity-Fabric/PCIe > 3X improvement in bandwidth!

USECASE: RESNET-50

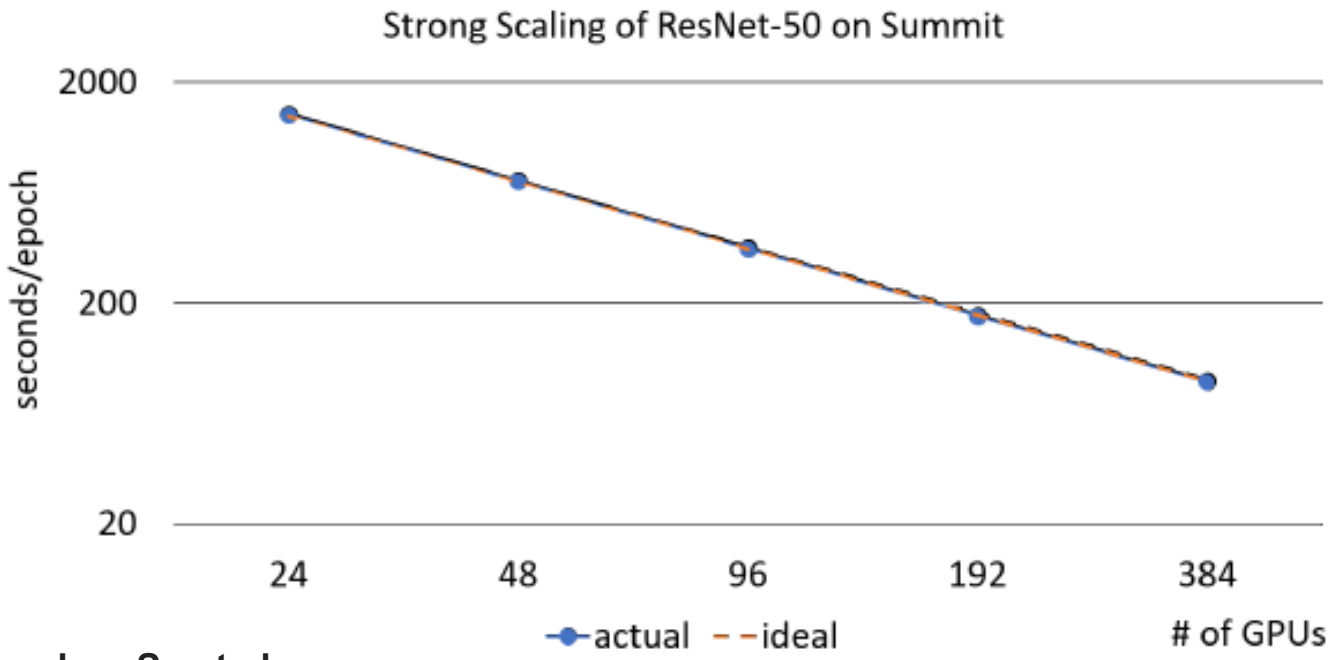


- Scale-up efficiency in excess of 95%
- Communication completely overlapped with computation
  - Except for first layer (very small impact)
- Batch size 32 – larger batch *reduces overall* comm time!

USECASE: RESNET-50 – WHAT HAPPENED

- Computation *12.5X the Non-blocking* Communication time
  - Gradients are 120MB in FP32
  - Latency negligible

RESNET-50 IS A BAD USE CASE



Vazhkudai, Sudharshan S., et al.
IEEE, 2018.

Fig. 14: Resnet-50 Scaling on Summit

AMD

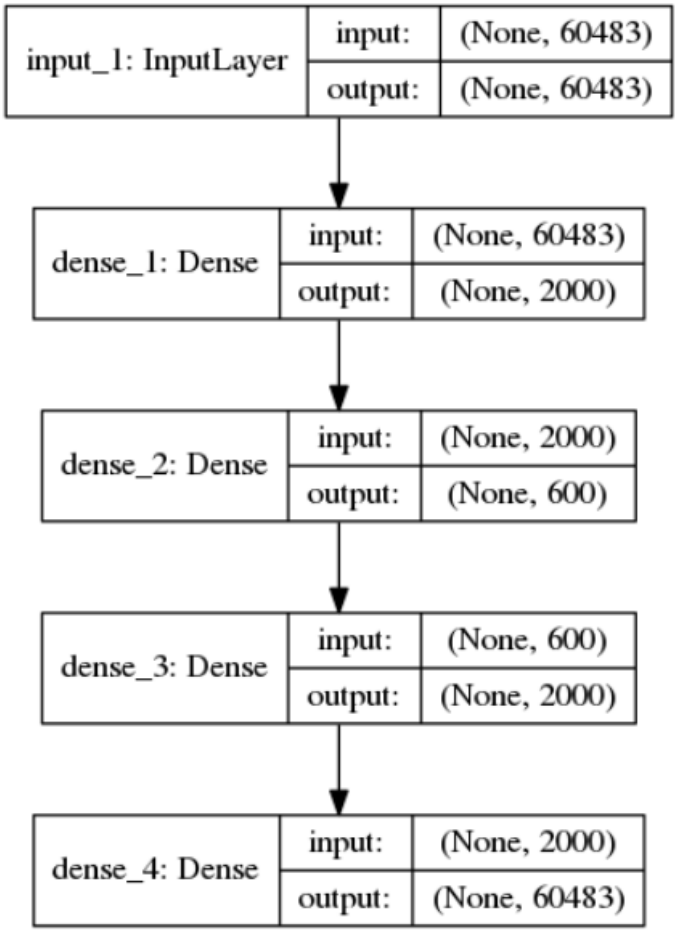
MY KINGDOM FOR A USE CASE

- That was bad use case!
  - Bad communication / computation ratio
  - Non-blocking communication

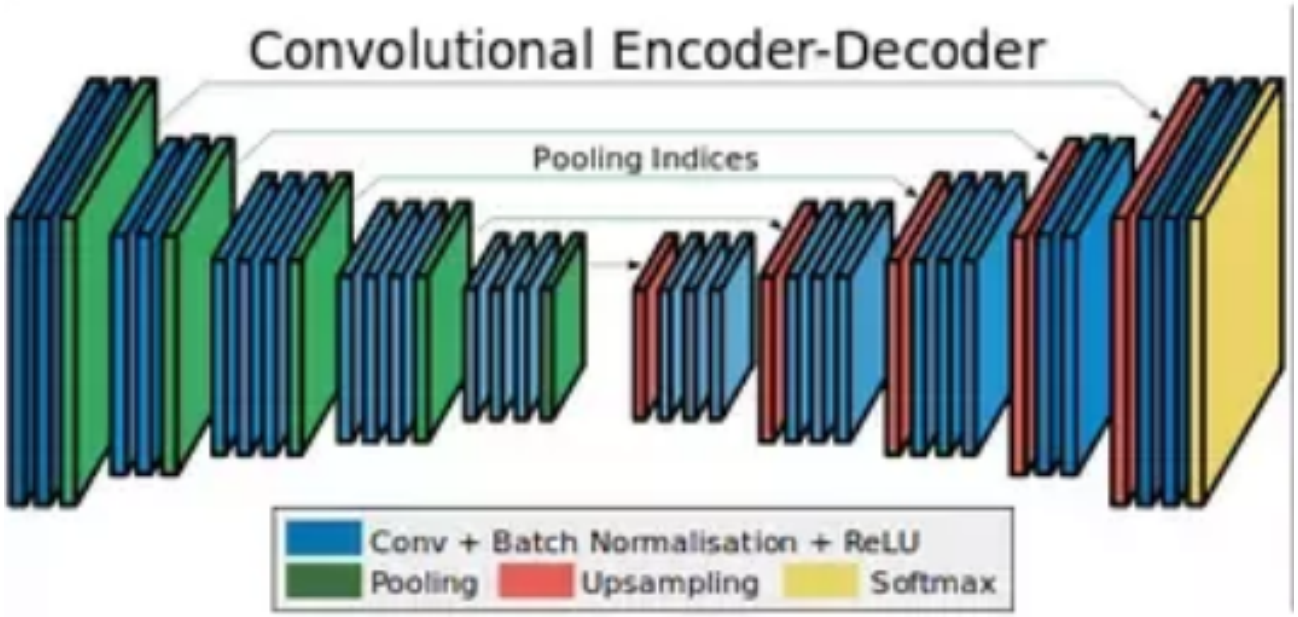
- What are the characteristics of good use cases?
  - High communication demands / computation
  - Blocking communication
  - ***Enormous* amounts of compute (Amdalh)**

USECASE: CANDLE P1B1

- **Autoencoder**
  - **Compressed Representation for Gene Expression**
  - Molecular models generate many features
  - Time-consuming (or impossible) to process
  - Also prone to over-fitting
  - Autoencoder:
    - use for feature reduction
    - input size and output size are identical



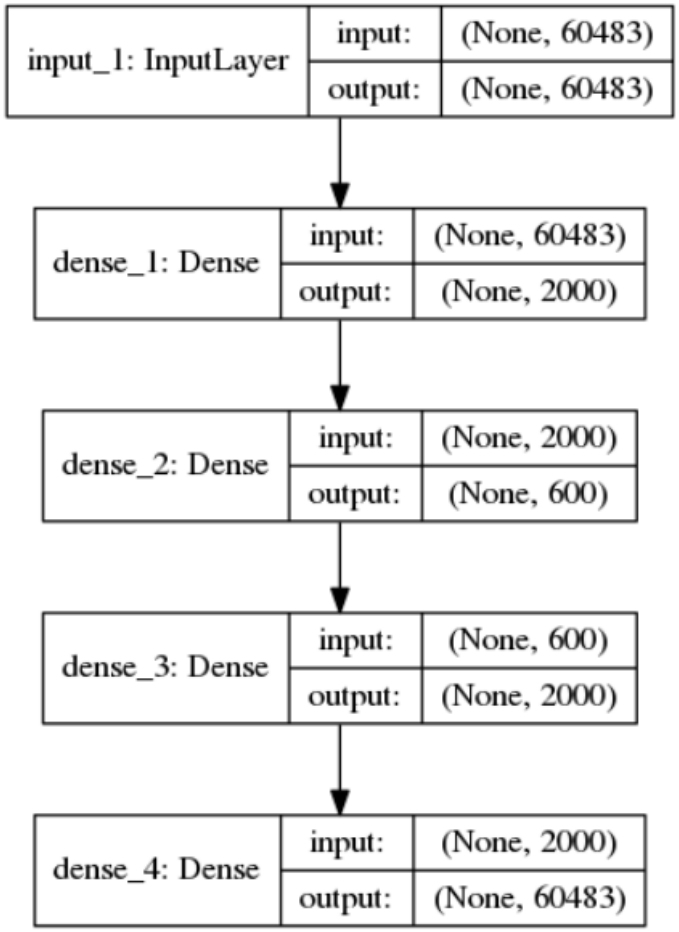
CANDLE P1B1: AUTOENCODER



- Uses DNN building blocks
- Hidden “Latent” vector

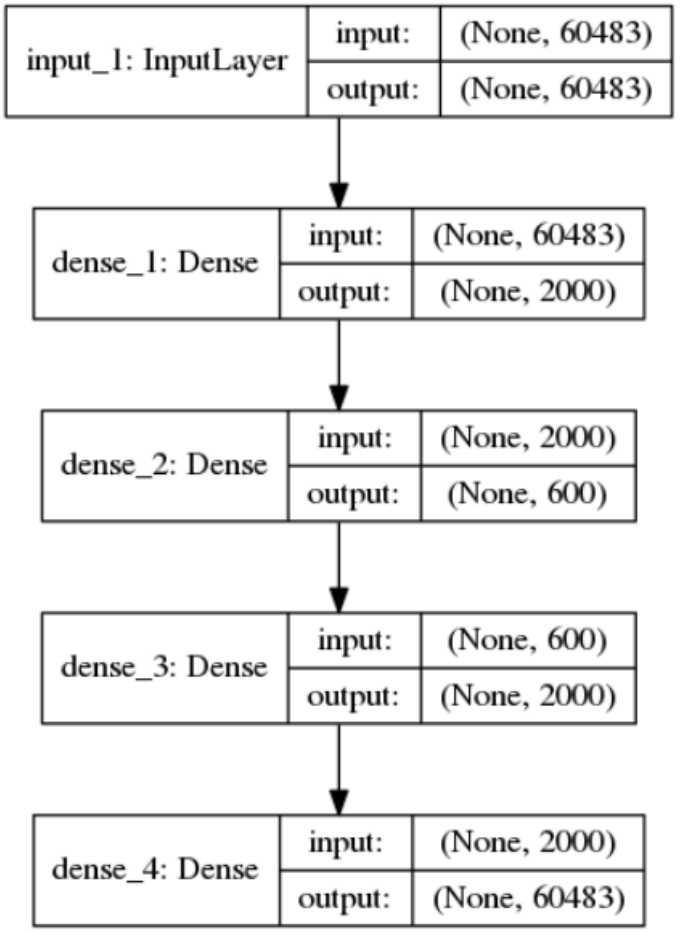
USECASE: CANDLE P1B1

- **Only 4 layers**
- **All dense (fully connected)**
- **Blocking Communication:**
  - Gradients in first layer are reduced across GPUs + applied to the weights before the next forward step



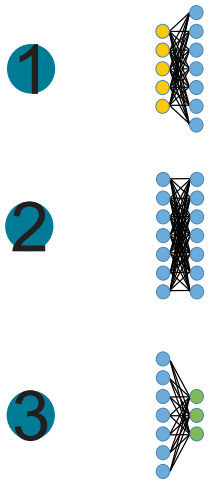
USECASE: CANDLE P1B1

- First layer input size: 60,483
  - Massive
- Total of 244 million parameters
- First and last layers have majority of parameters
  - ~120 million each
- "Data transfer of 480Mbytes/GPU for the layer that cannot be overlapped with any computation"

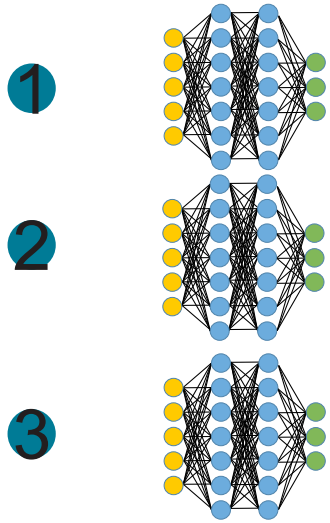


CIRCLING BACK: PARALLELISM TYPES

WHICH TYPES OF DL PARALLELISM MAKES SENSE? AND WHEN?



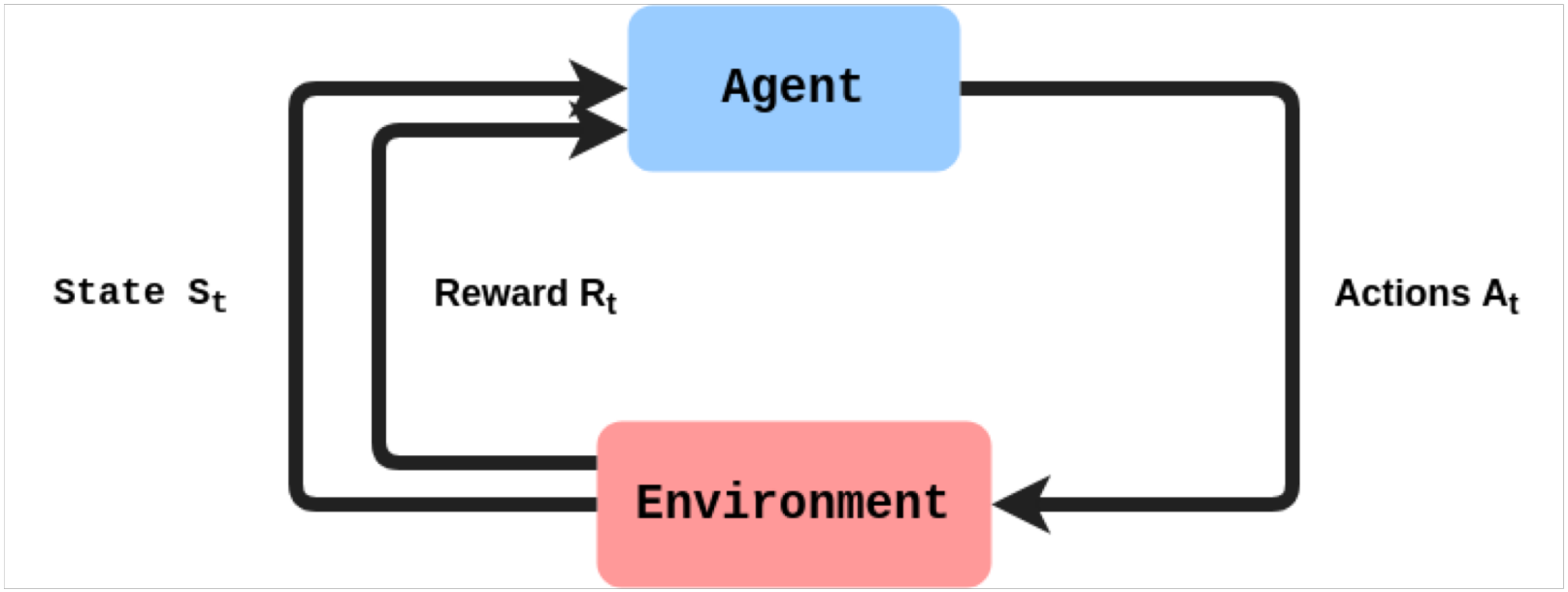
Model Parallelism



Data Parallelism

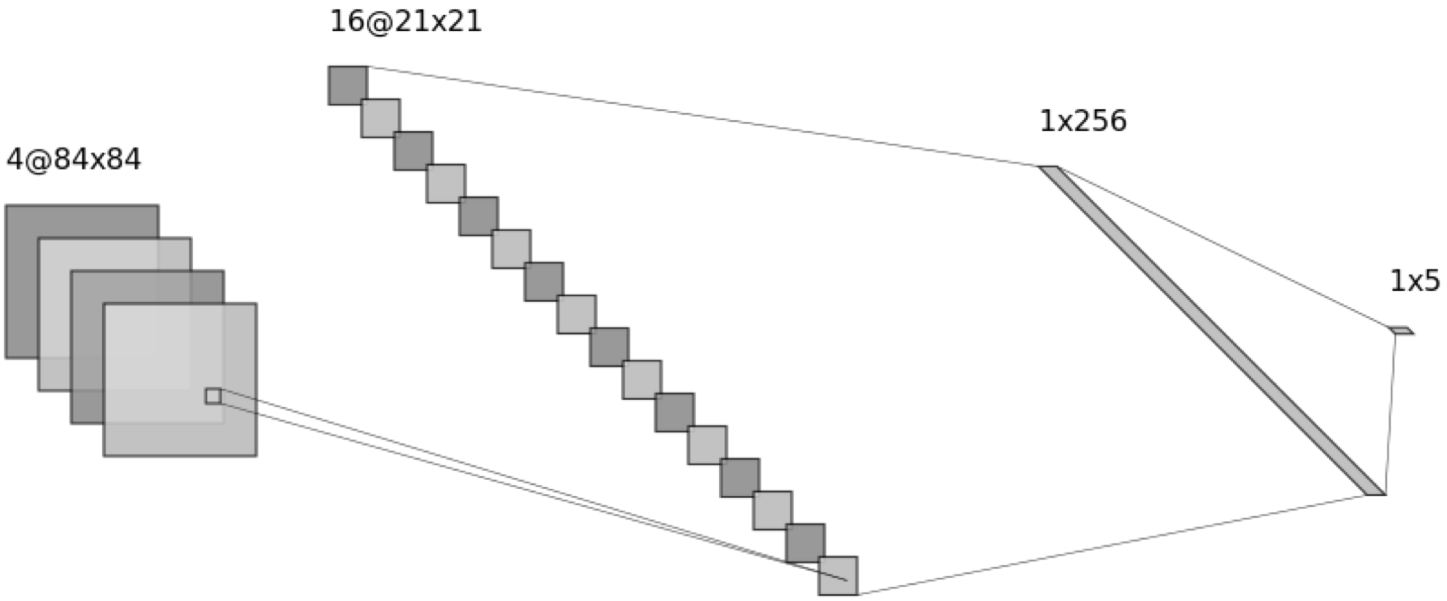
REINFORCEMENT LEARNING

REINFORCEMENT LEARNING



DEEP CAN BE MISLEADING

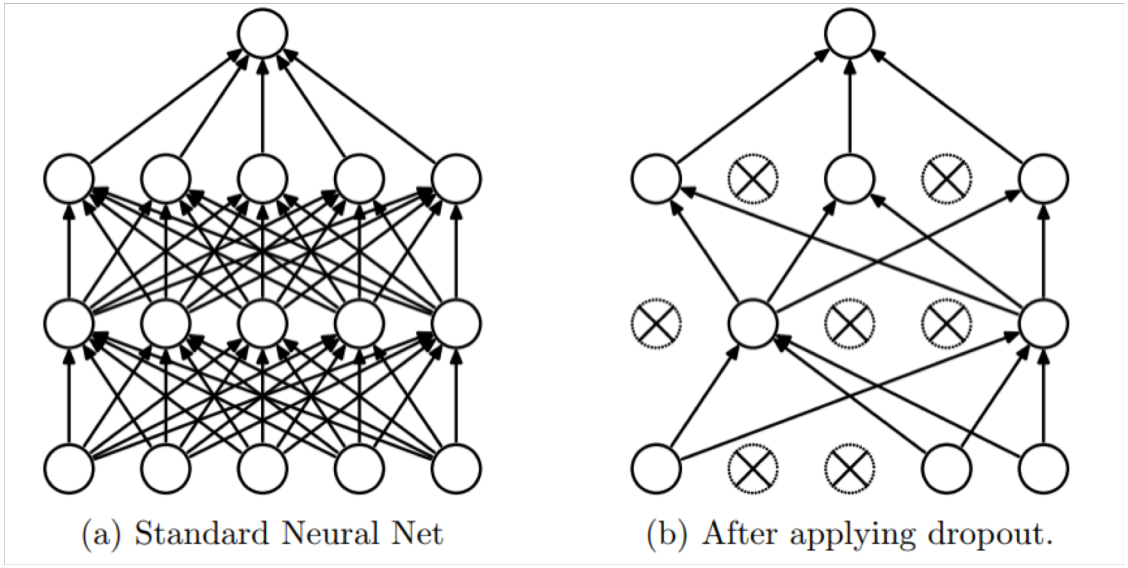
•Not all state-of-the-art networks are high capacity (Atari DQN)



OVERFIT

Joint work with: Jiajing Guan, Patrik Gerber, Elvis Nunez, Kaman Phamdo

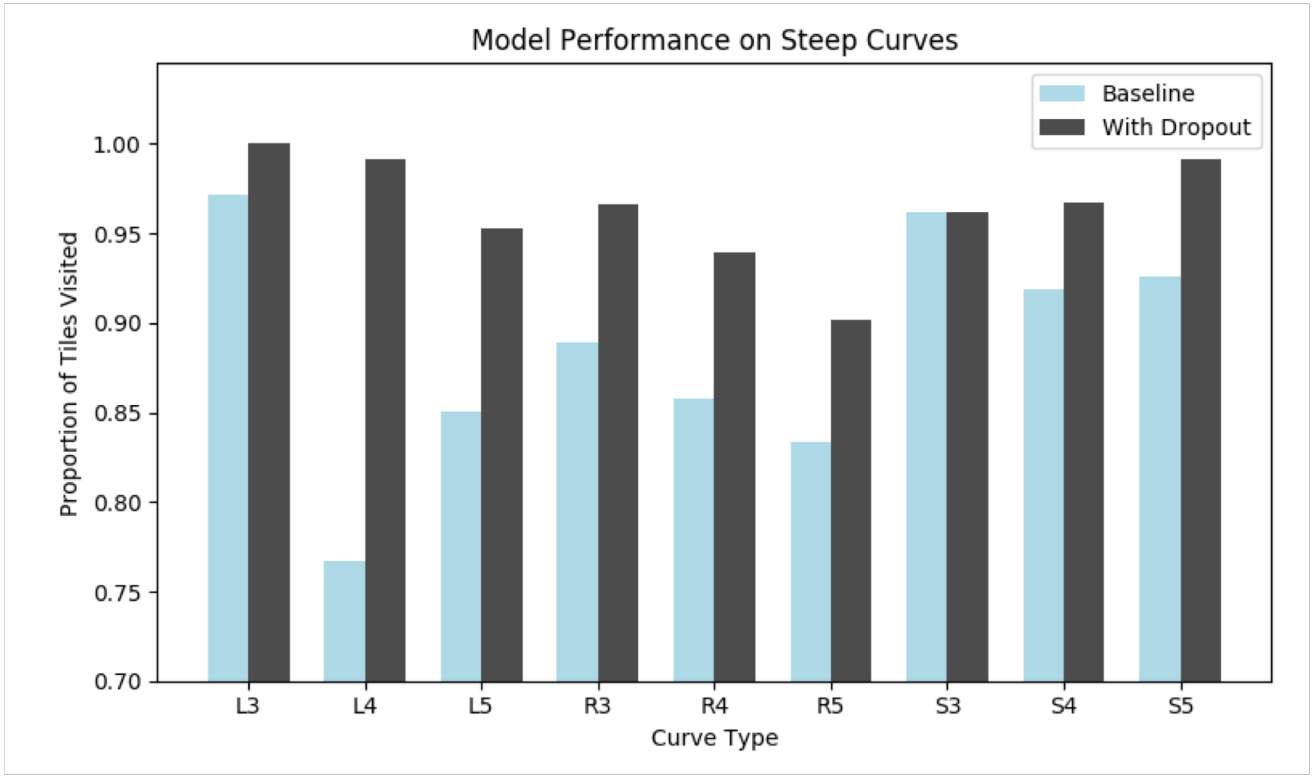
•Driving towards higher capacity



REINFORCEMENT LEARNING

https://drive.google.com/file/d/1DQU4yCsq6nbVJB6WKoXlED9YFGDsellu/view

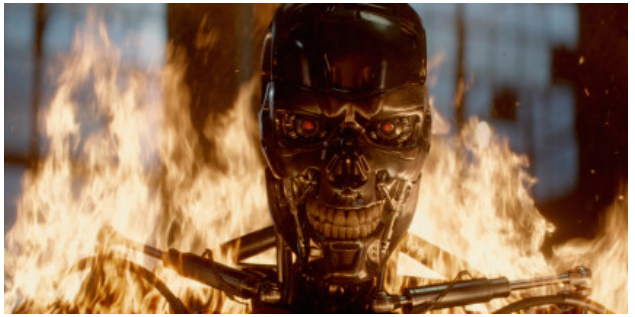
REINFORCEMENT LEARNING



AMD

CONCLUSIONS

WHAT MACHINE LEARNING CANNOT DO



- Garbage in/out:
  - Military Classifier: Tank vs. Car
  - Poison attacks
  - Adversarial Attacks

Ask Alexa/Siri, etc. a non-trivial question

Challenging to debug / sanity check

Not taking the job of anyone in this room!

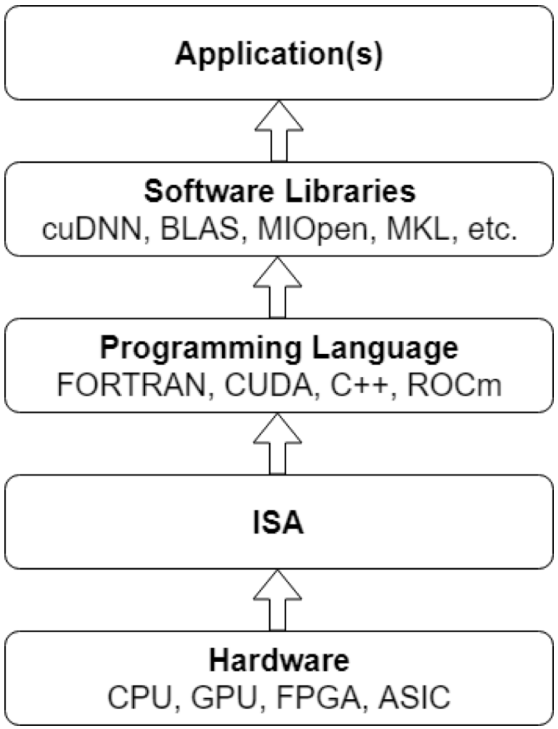
Do not expect a neural net to generate your final report

CONCLUSIONS

- This is not a singular field
  - Different approaches necessary
  - No “one hardware to rule them all”
  - “Cambrian explosion” of hardware is ongoing

- Important Research Challenges:
  - System-level design (sell capabilities, not hardware)
  - Theory (does not have to be alchemy)
  - Interdisciplinary: look at HPC, Physics, Neuroscience

MORE THAN CAT PHOTOS



- Full-Stack transparency necessary for science

ACKNOWLEDGEMENTS

- Huge number of contributors across AMD
  - Research
  - RTG (Radeon Technology Group)
  - CPU/Servers
  - Edge Inference Team

  - And many external collaborators in Academia, Industry, Government, etc.

JOIN US!

- Opportunities in:
  - Research
  - Machine Learning Software Libraries
  - Edge Inference
  - Hardware Architecture
  - COOPs, Internships, etc.

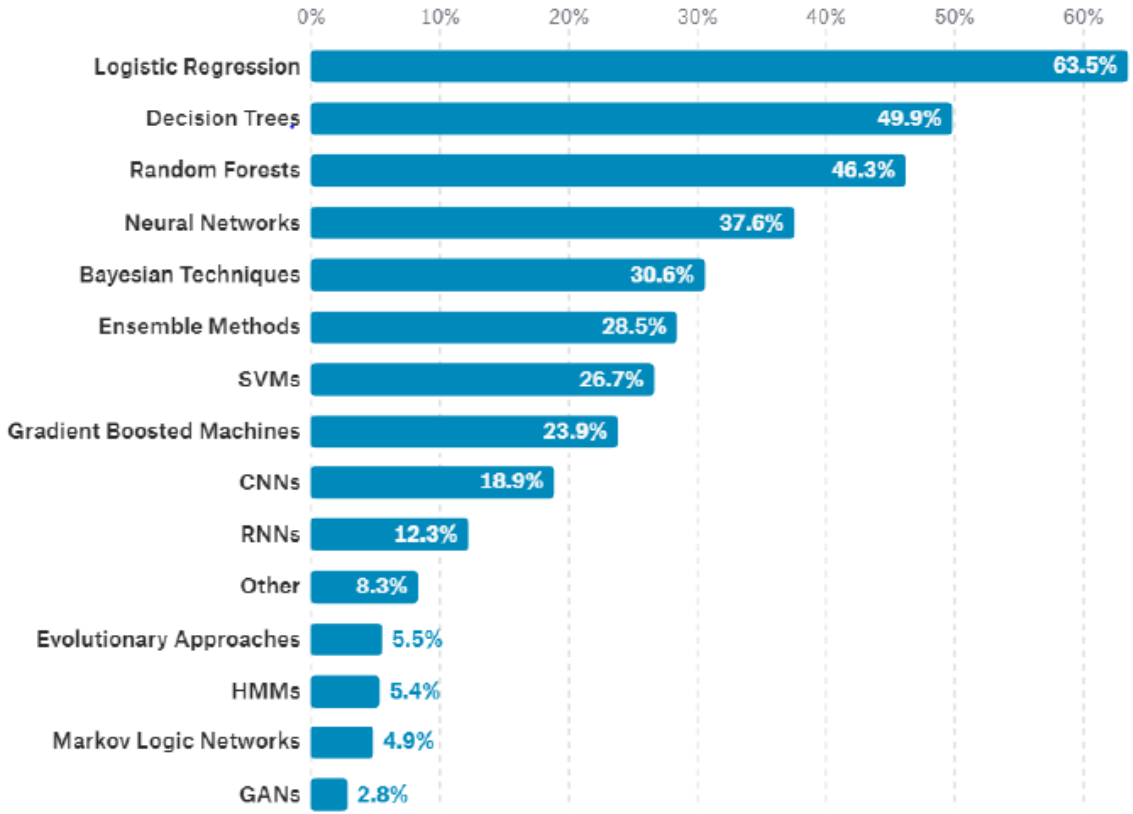
- **We have openings in Spring**

QUESTIONS?

nicholas.malaya@amd.com

BACKUP SLIDES

CNNS/GANS, ETC., NOT WIDELY DEPLOYED (YET)



AMD

SHORTCOMINGS

- Who is liable?
- How to certify these systems?