Deep Learning: Convergence of HPC and Hyperscale

Sumit Sanyal
CEO
Technology Revolution: Deep Neural Nets

- A new paradigm in programming
  - DNNs remarkably effective at tackling many problems
  - Designing new NN architectures as opposed to “programming”
  - Training as opposed to “compiling”
  - Trained weights as the “new” binaries

- Big Neural Nets required to process Big Data
  - Videos, images, speech and text
  - DNNs are significantly increasing recognition accuracies which have stagnated for decades
  - Used to structure Big Data
Neural network development

1. Make a guess at NN architecture
2. Design bigger network
3. Train on labeled dataset
4. Run on training data, does it perform well?
   - Yes: Move to next step
   - No: Iterate
5. Run on test data, does it perform well?
   - Yes: DONE!
   - No: Iterate
6. Need more training data
7. Iterate
Deep Neural Nets require a lot of compute cycles!

An example – image classification:
- Ratio of (Compute cycles : IO bandwidth) significantly higher than non AI algorithms
- Training AlexNet (for image classification) requires ~27,000 flops/input data byte
- Training VGG ~150,000 flops/data byte

$R^3 / R^2 \rightarrow$ Volume (compute) / Surface(IO BW)
- Significantly higher for Deep Nets

Power dissipation challenges
- Compute density limited by DC cooling capacity
- At 1 ~uW/ MHz (current state-of-art in 28 nm) requires 300 Watts!

AI is no longer bored 😊
Neural Net Computations

- All Deep Neural Net implementations have the following properties
  - Small set of non-linear transforms
  - Small set of linear algebra primitives
  - Relatively modest dynamic range of weight/data values
  - Very regular/repetitive data flows
  - Only persistent memory requirement is for weights
    - Updated while learning, fixed for recognition

- Variance in the size of the net across applications is $>10^5$

Compute cycles will be commoditized; not computers!
Why minds.ai?

“…around 2008 my group at Stanford started advocating shifting deep learning to GPUs (this was really controversial at that time; but now everyone does it); and I'm now advocating shifting to HPC (High Performance Computing/Supercomputing) tactics for scaling up deep learning. Machine learning should embrace HPC. These methods will make researchers more efficient and help accelerate the progress of our whole field.”  – Andrew Ng, Feb 2016

minds.ai was founded on the basis that Deep Learning is the next frontier for HPC – it’s what we do
7 levels of parallelism

- Instruction level – SIMD, VLIW etc.
- Thread level – warps
- Processor level – many cores
- Server level – many GPUs in a server
- Cluster level – many servers with high BW interconnect
- Data Center level
- Planet level 😊
ASICS for DNNs

- Exponential growth in Data Centers
  - Commoditization of Enterprise silicon
  - Traditional mobile players announcing ASICS for enterprise compute

- Higher demands for compute density
  - GPGPUs have won the first round
  - Dennardian scaling is breaking down

- Power dissipation will emerge as major challenge
  - Chip level, server level and DC level
### Age of dark silicon

<table>
<thead>
<tr>
<th>Transistor property</th>
<th>Dennardian</th>
<th>Post Dennardian</th>
</tr>
</thead>
<tbody>
<tr>
<td># of transistors (Q)</td>
<td>$S^2$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Peak clock frequency (F)</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>Capacitance (C)</td>
<td>$1/S$</td>
<td>$1/S$</td>
</tr>
<tr>
<td>Supply Voltage (V_{dd})</td>
<td>$1 / S^2$</td>
<td>$1$</td>
</tr>
<tr>
<td>Dynamic Power (QFCV_{dd}^2)</td>
<td>$1$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Active Silicon</td>
<td>$1$</td>
<td>$1/S^2$</td>
</tr>
</tbody>
</table>

* $S$ is the ratio of feature size between next generation processes
Heterogeneity in Enterprise silicon

- Dark silicon will drive heterogeneity
  - Multi core architectures with different Instruction Sets
  - Power aware scheduling across cores
  - Decreasing parts of the chip can run at full clock frequency

- Specialized silicon for server blades
  - Bridges to intra server and inter server communications
  - Last level caching support, caching across MPI
  - Distributed compute in network interfaces
7 levels of parallelism

- Instruction level – SIMD, VLIW etc.
- Thread level – warps
- Processor level – many cores
- Server level – many GPUs in a server
- Cluster level – many servers with high BW interconnect
- Data Center level
- Planet level 😊
Deep Learning is driving the convergence of High Performance Compute (HPC) and Hyperscale (Data Centers)
- Traditional HPC ecosystems: expensive and bleeding edge
- DC infrastructure: commodity and homogeneous
  - Single or dual CPU servers common

All of this is changing
- GPGPUs now common in DCs, initial resistance
- InfiniBand penetration has reached a tipping point
- Dense compute clusters require high bandwidth interconnects
Server Architectures

- Intra server vs. Inter server bandwidths
  - Inter server bandwidths will grow faster than intra server
  - Lead to larger, denser servers
    - 8 or more GPGPUs per server for DNN training jobs
    - Will co-exist with CPU based servers for search and database operations
  - Many kinds of servers, one size fits all does not work
Training Server Reference Design
• Scalable up to 7 Server Nodes
• 8 GPUs per Node
• ~2.5kw per Server Node
• 100Gbps IB Interconnect
7 levels of parallelism

- Instruction level – SIMD, VLIW etc.
- Thread level – warps
- Processor level – many cores
- Server level – many GPUs in a server
- Cluster level – many servers with high BW interconnect
- Data Center level
- Planet level 😊
Data Center Architectures

- Computing super clusters
  - $10^5$ variability in size of compute ‘jobs’
  - Large number of collocated servers running the same job
  - High BW, low latency interconnect
  - Clusters could grow to significant fraction of a Data Center

- Architecture of clusters will be hierarchical and heterogeneous
  - Edge servers for security and Data management
  - Dedicated RAID servers
  - Dedicated compute servers
  - Control and management nodes

- Multi DC training clusters for big models are technically feasible
  - Scientific community has performed trans continental simulations
Thank you.
Contacts

• Sumit Sanyal, Founder and CEO
  sumit@minds.ai

• Steve Kuo, Co-Founder
  steve@minds.ai
minds.ai
Accelerated Deep Learning