# DjiNN and Tonic: DNN as a Service and Its Implications for Future Warehouse Scale Computers

**Johann Hauswald**, Yiping Kang, Michael A. Laurenzano, Quan Chen, Cheng Li, Trevor Mudge, Ronald G. Dreslinski, Jason Mars, Lingjia Tang

University of Michigan — Ann Arbor, MI

DjiNN
& Tonic

COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

ClarityLab

# Intelligent Personal Assistant (IPA) Queries

# Intelligent Personal Assistant (IPA) Queries

“Set my alarm for 6am.”

# Intelligent Personal Assistant (IPA) Queries

# Intelligent Personal Assistant (IPA) Queries

“What’s the
speed of light?”

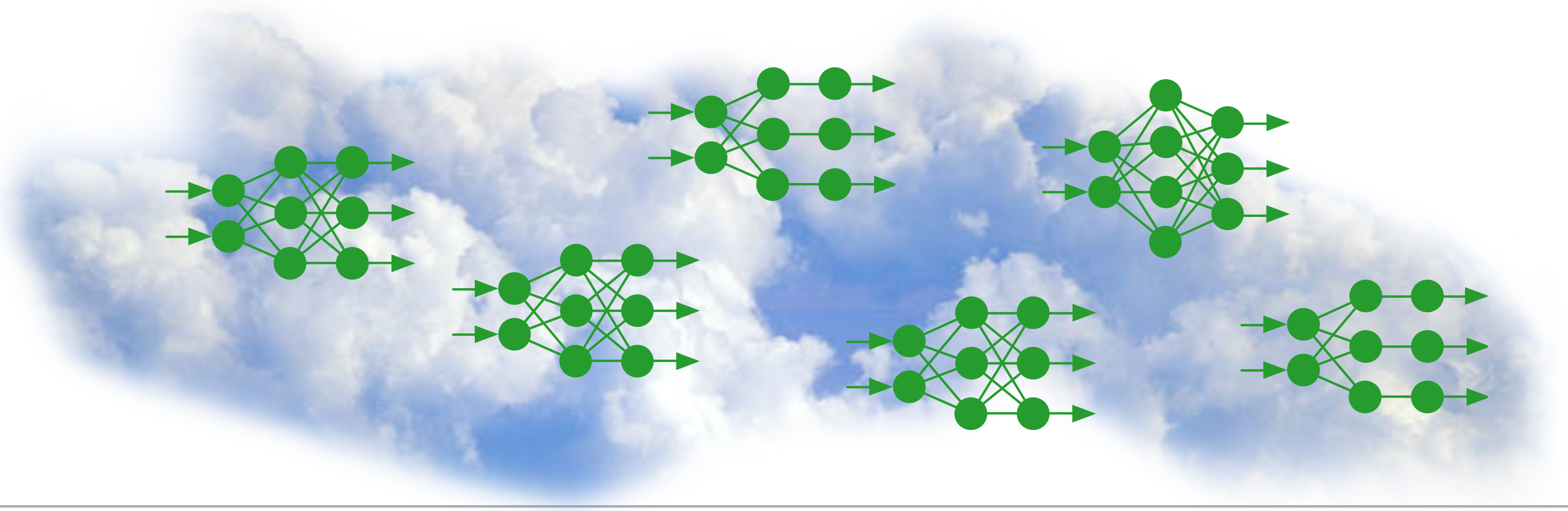
# Intelligent Personal Assistant (IPA) Queries

# Intelligent Personal Assistant (IPA) Queries

“Who is this?”

# Intelligent Personal Assistant (IPA) Queries

# Intelligent Personal Assistant (IPA) Queries



ROBERT MCMILLAN 06.30.14 6:30 AM

SIRI WILL SOON UNDERSTAND YOU A WHOLE LOT BETTER

Forbes / Tech

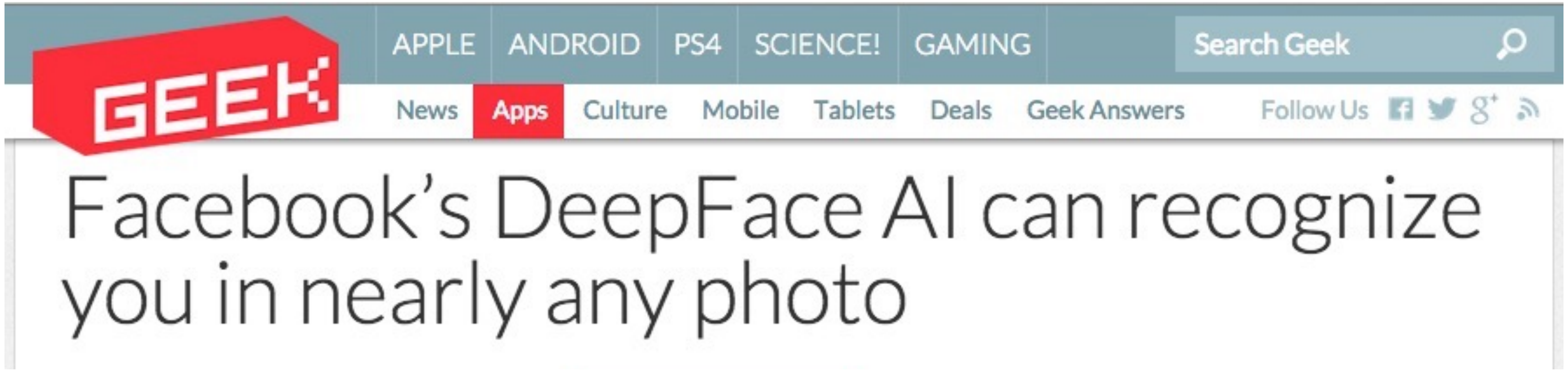
FEB 19, 2015 @ 1:06 PM 6,203 VIEWS

Microsoft's Deep Learning Project Outperforms Humans In Image Recognition

The Register®

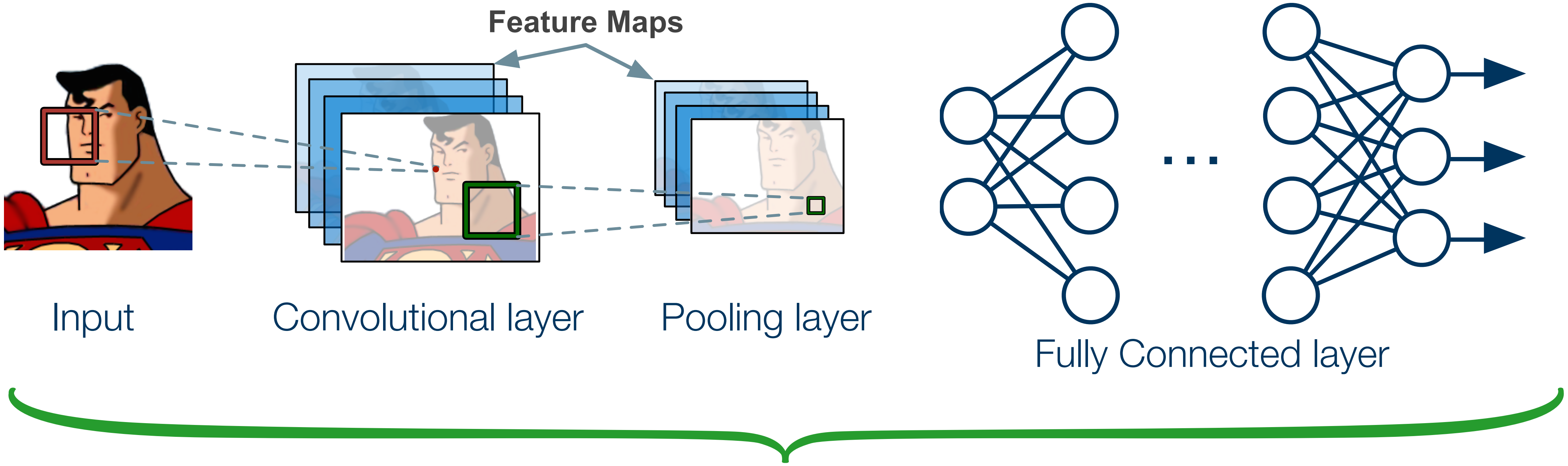
Biting the hand that feeds IT

IBM buys 'deep learning' upstart Alchemy API, inhales 40,000 devs



Facebook's DeepFace AI can recognize you in nearly any photo

# Deep Neural Networks (DNNs)



Feature Maps

Input

Convolutional layer

Pooling layer

Fully Connected layer

Network Architecture

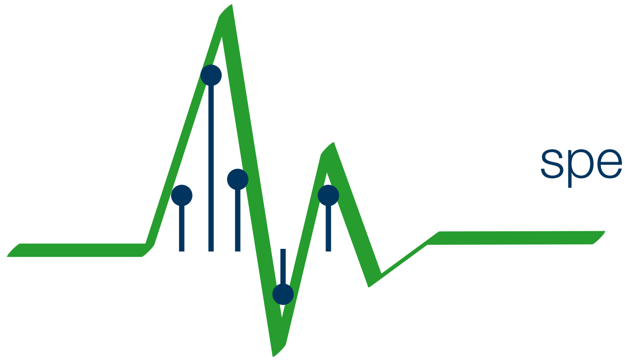
Inference

0.1 “Spiderman”

**0.9 “Superman”**

0.5 “Batman”

# Deep Neural Networks (DNNs)



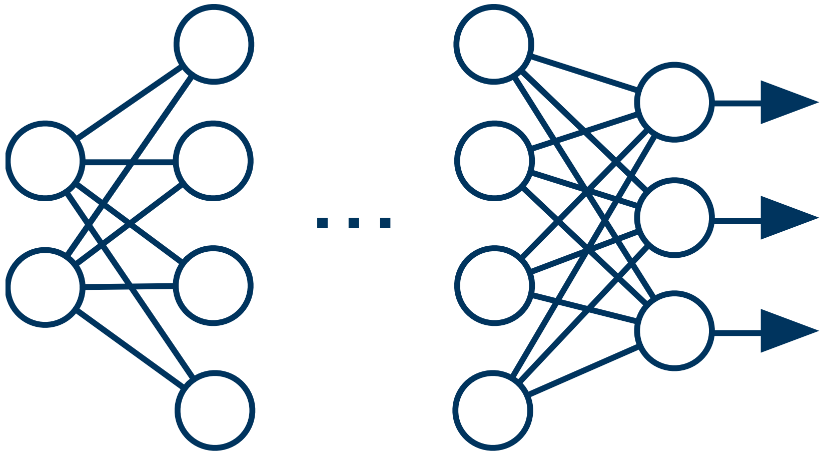
speech features

word vectors

“who” → $[w_0\, w_1 \ldots w_k]$

“is” → $[w_0\, w_1 \ldots w_k]$

“this” → $[w_0\, w_1 \ldots w_k]$



Fully Connected layer

Network Architecture

Inference

“Who”, “is”, “this”

“Who” (PRONOUN)
“is” (VERB)
“this” (PRONOUN)

# DNN as a Service

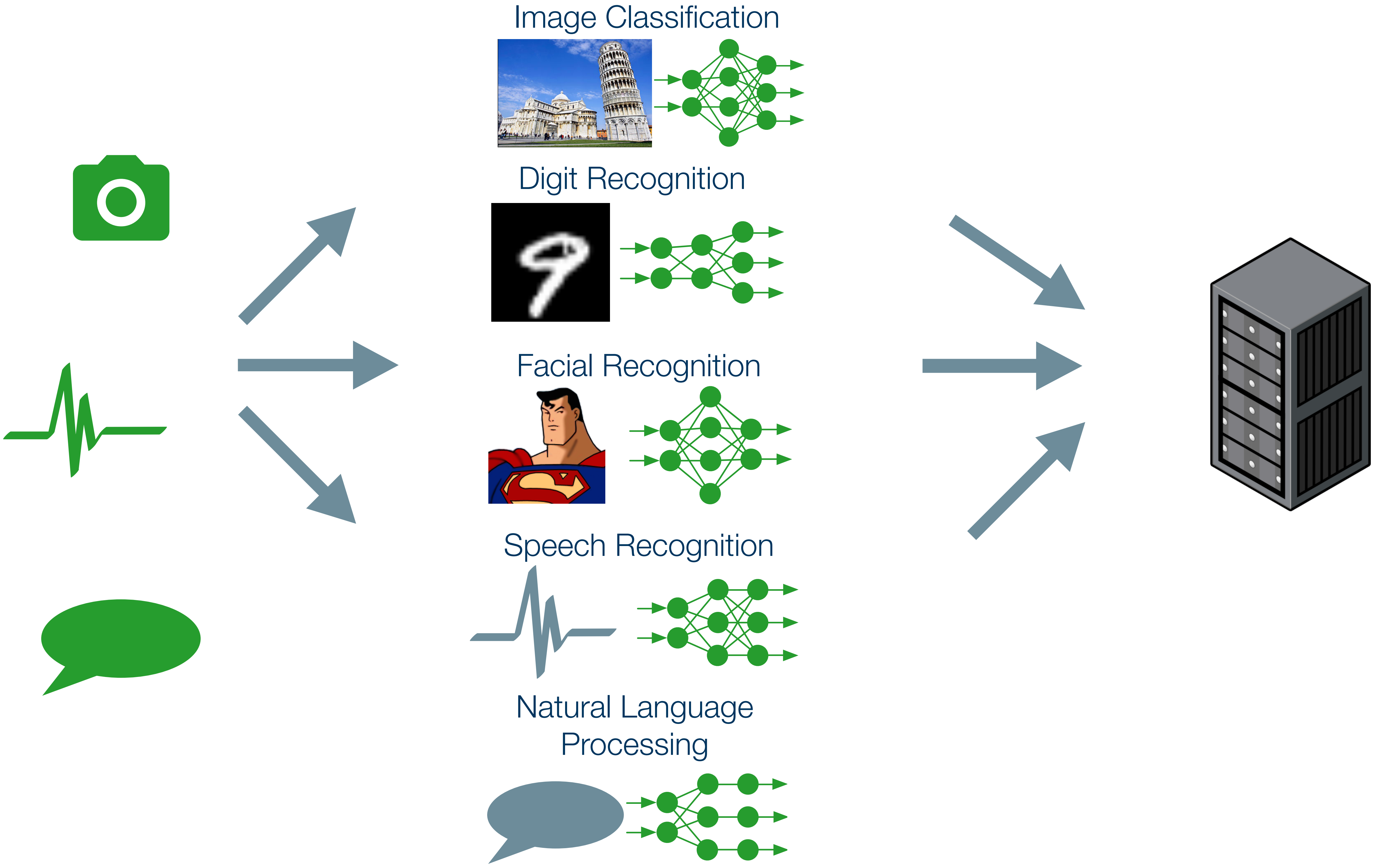


Image Classification

Digit Recognition

Facial Recognition

Speech Recognition

Natural Language Processing

# DNN as a Service

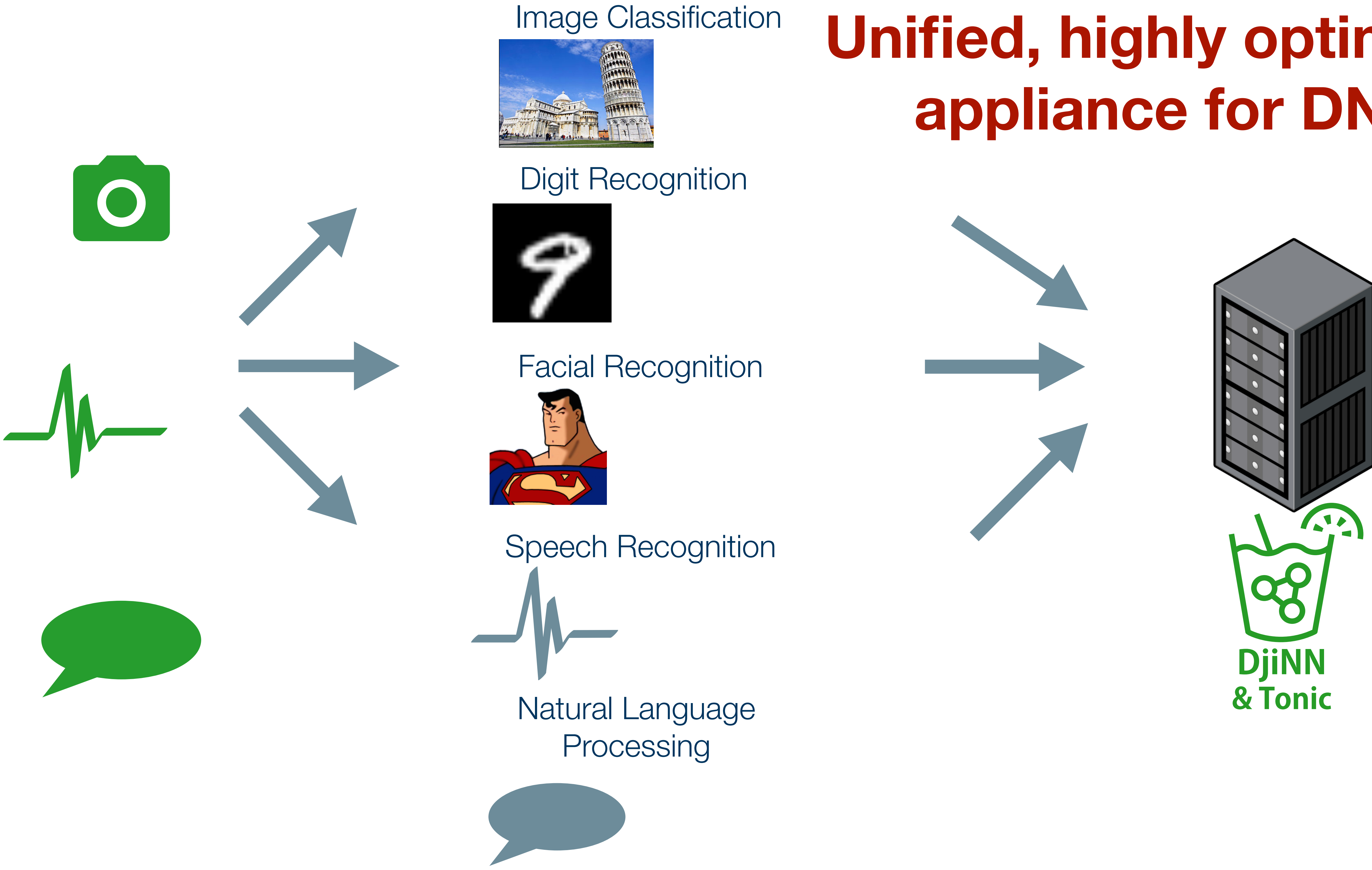


Image Classification

Digit Recognition

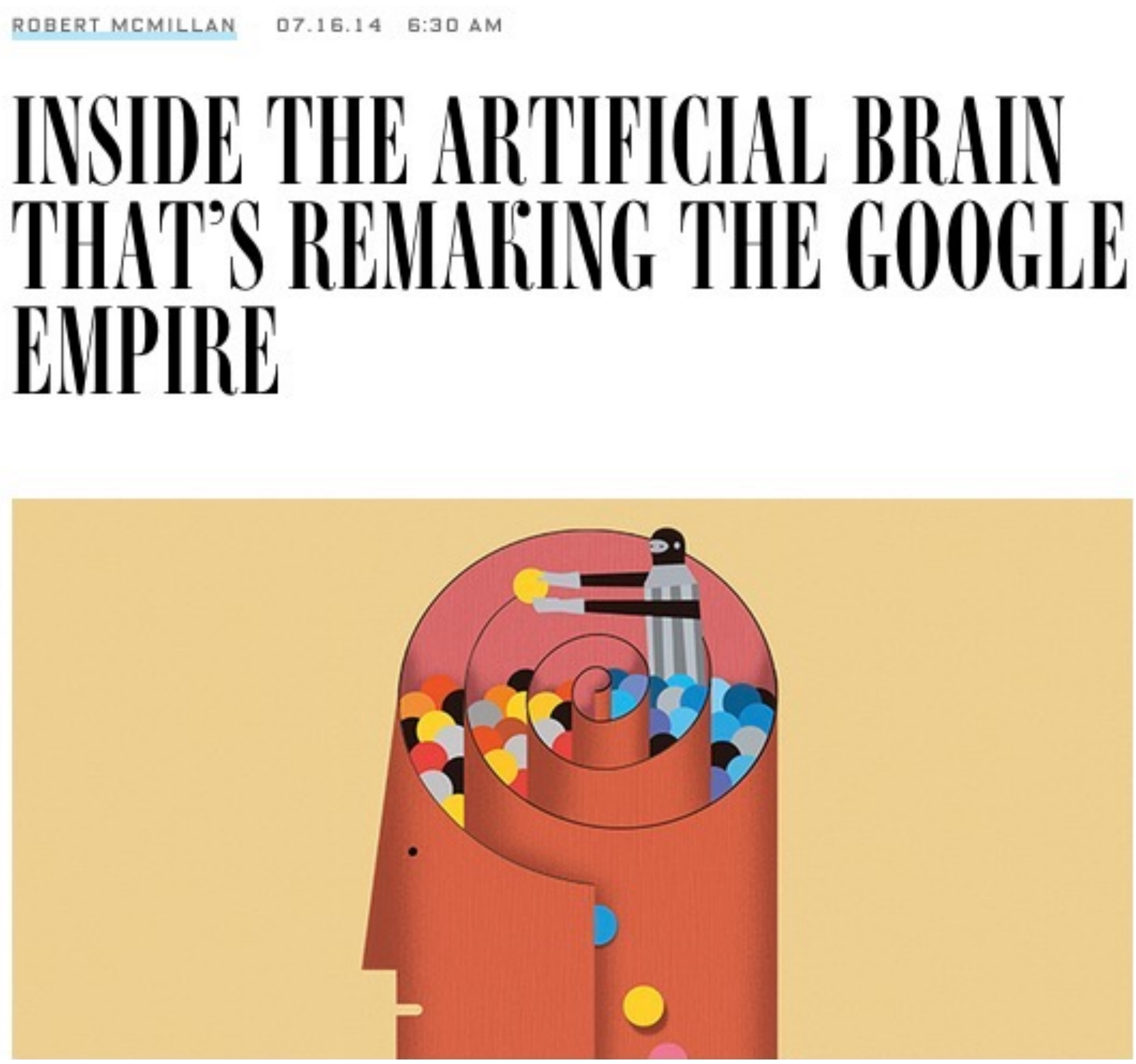
Facial Recognition

Speech Recognition

Natural Language Processing

DjiNN & Tonic

**Unified, highly optimized appliance for DNN**



ROBERT MCMILLAN 07.16.14 6:30 AM

# INSIDE THE ARTIFICIAL BRAIN THAT'S REMAKING THE GOOGLE EMPIRE

*“We now have probably 30 or 40 different teams at Google using our infrastructure.”*

-Jeff Dean

# DNN as a Service

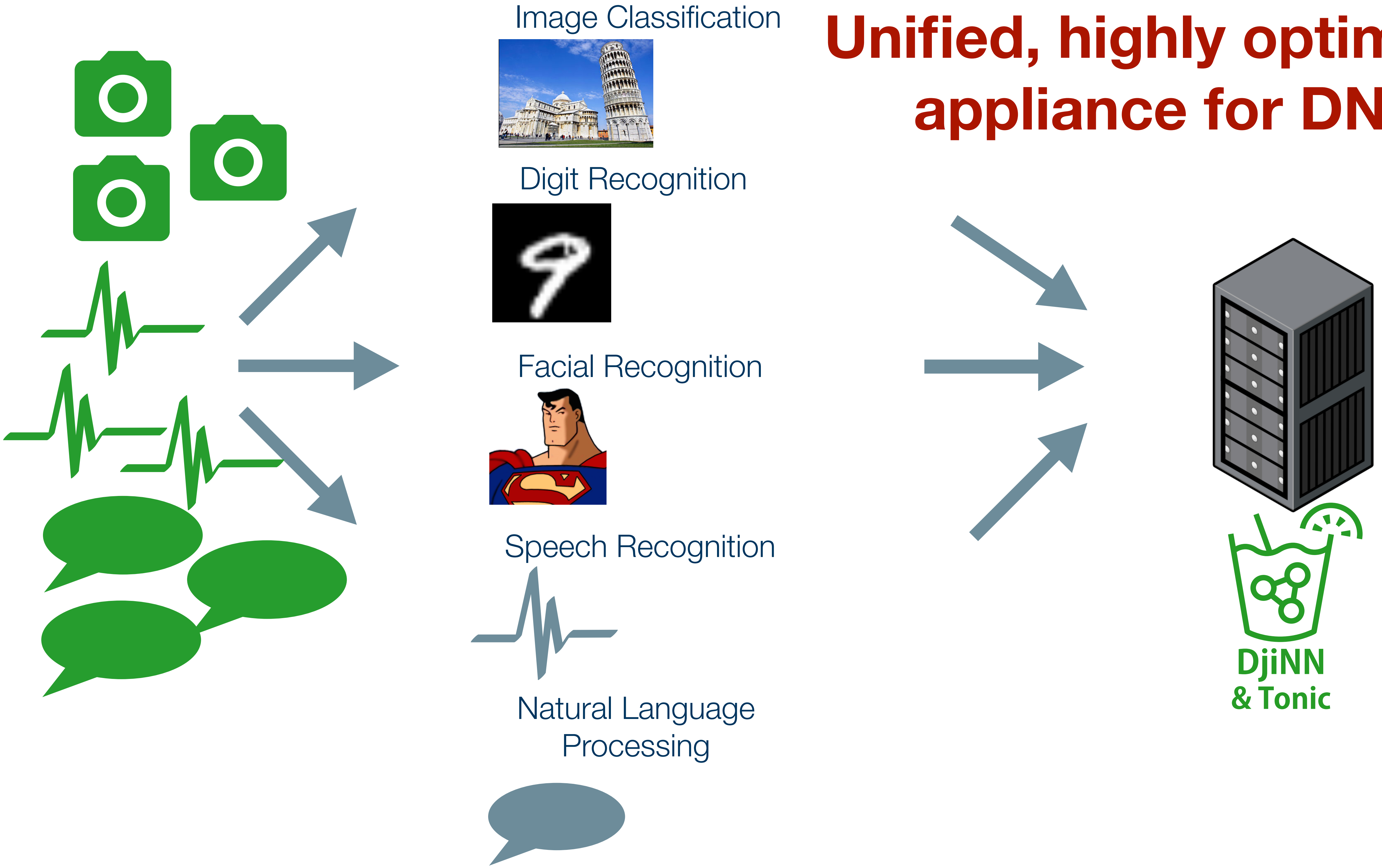


Image Classification

Digit Recognition

Facial Recognition

Speech Recognition

Natural Language Processing

**Unified, highly optimized appliance for DNN**

# Challenge

Design a high throughput Warehouse Scale Computer (WSC)
for DNN as a Service

# Outline

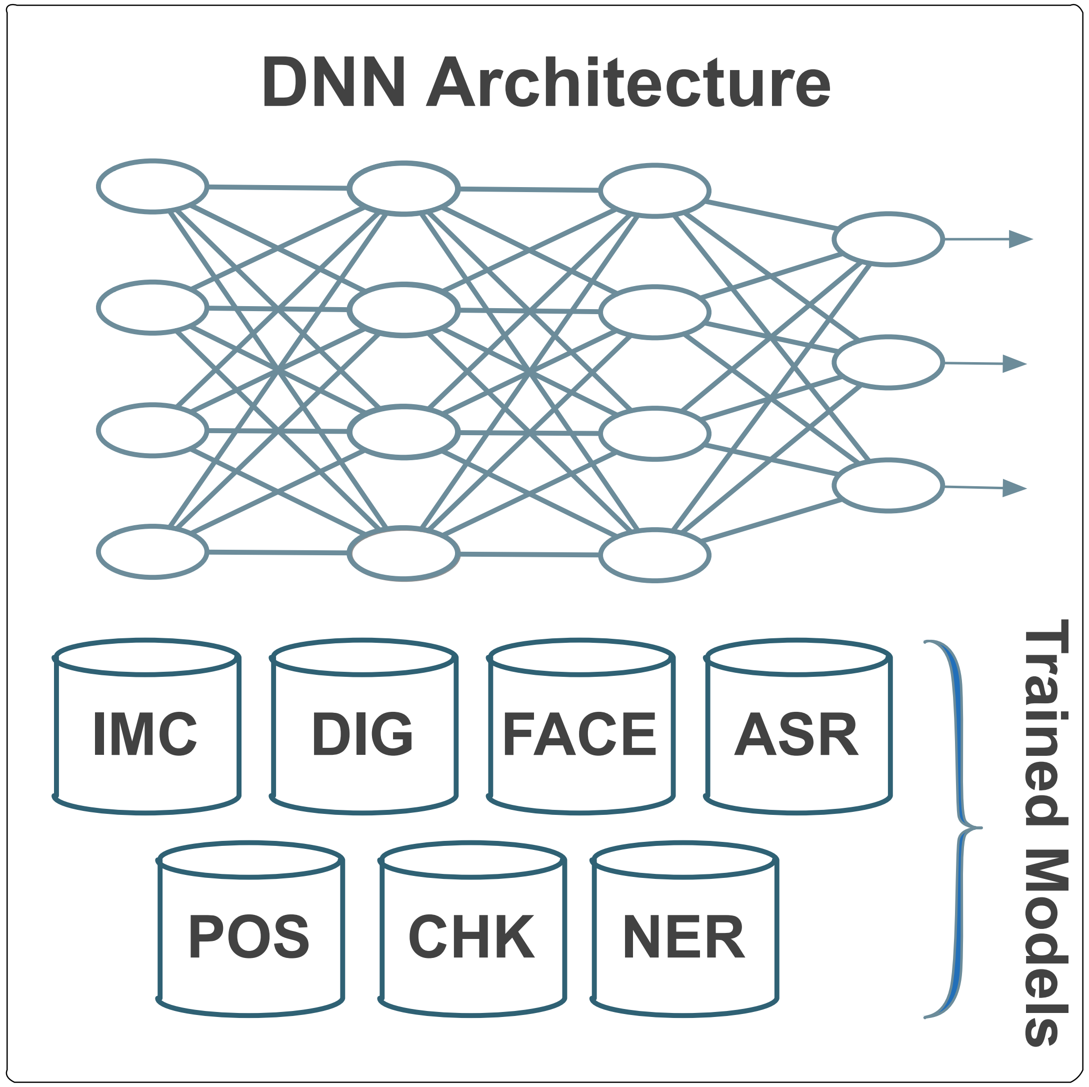
- DjiNN and Tonic: DNN as a Service
- Identifying Bottlenecks for DNN as a Service
- Designing a High Throughput System
- Future Warehouse Scale Computer (WSC) Designs
- Conclusion

# DjiNN and Tonic: DNN as a Service

# DjiNN Design Goals

- Single web-service for DNN
- Diverse applications
- Low overhead request processing

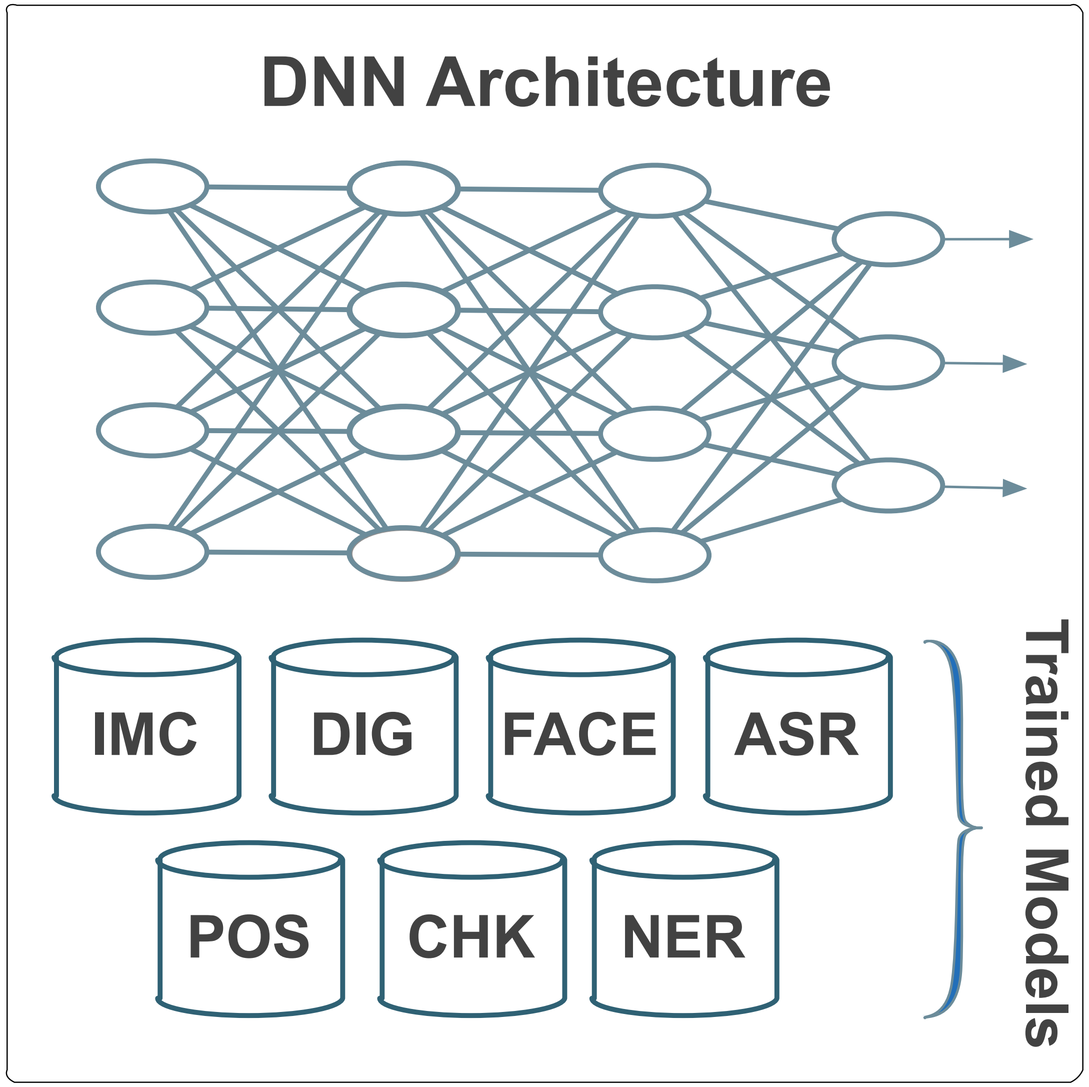
***DjiNN DNN Service***



# DjiNN Implementation

- Decoupled architecture
- Arbitrary network architecture support
- Memory resident models for thread pool
- More details in paper

***DjiNN DNN Service***



# Tonic Suite

- End-to-end applications that make requests to the DjiNN Service
- Span image, speech, and natural language processing
- State-of-the-art neural network architectures

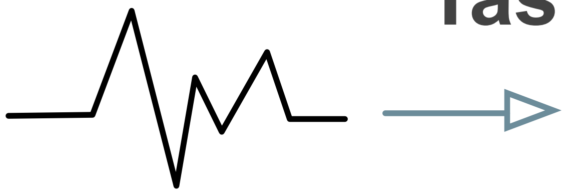
**Image Task**

**IMC**

**DIG**

**FACE**

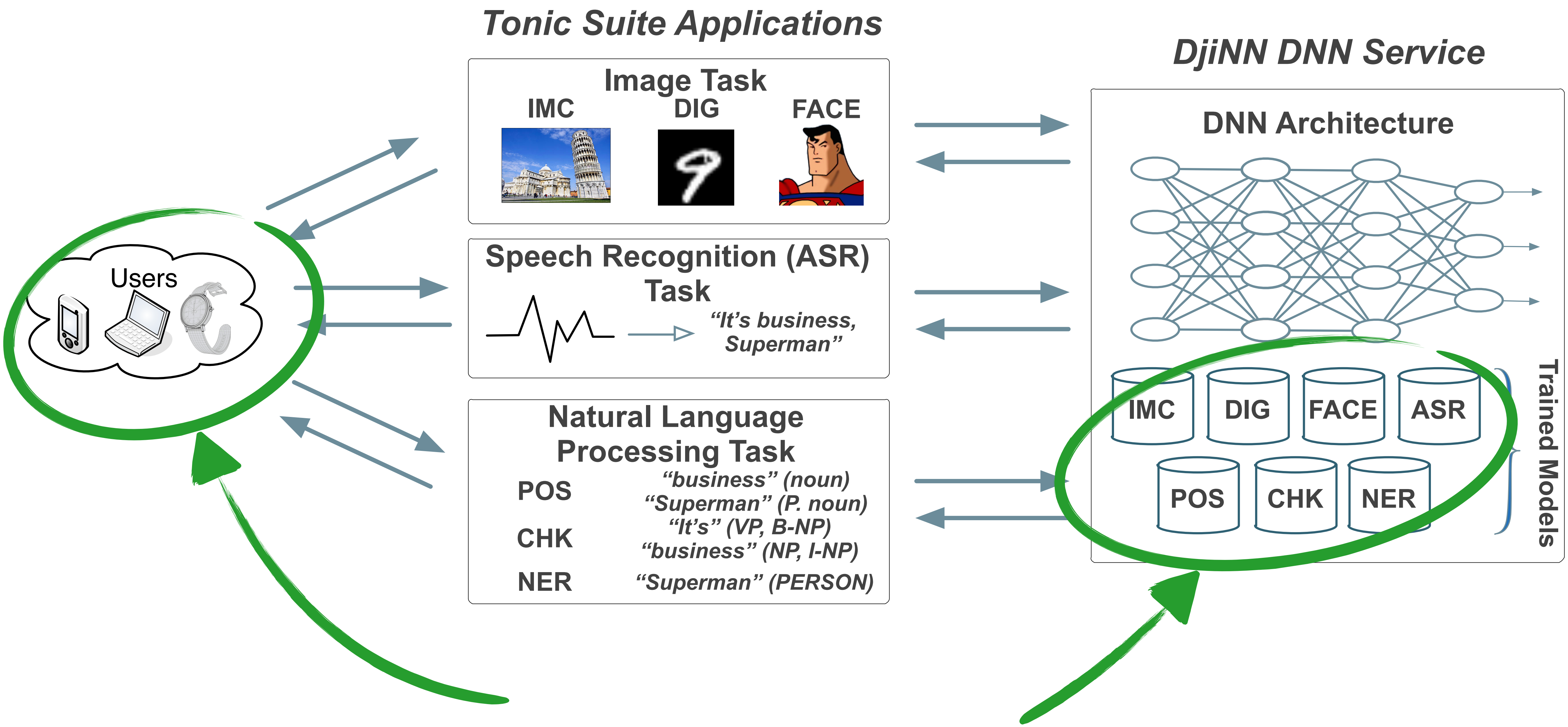
**Speech Recognition (ASR) Task**



***"It's business, Superman"***

**Natural Language Processing Task**

| | |
|---|---|
| **POS** | ***"business" (noun)***<br>***"Superman" (P. noun)*** |
| **CHK** | ***"It's" (VP, B-NP)***<br>***"business" (NP, I-NP)*** |
| **NER** | ***"Superman" (PERSON)*** |



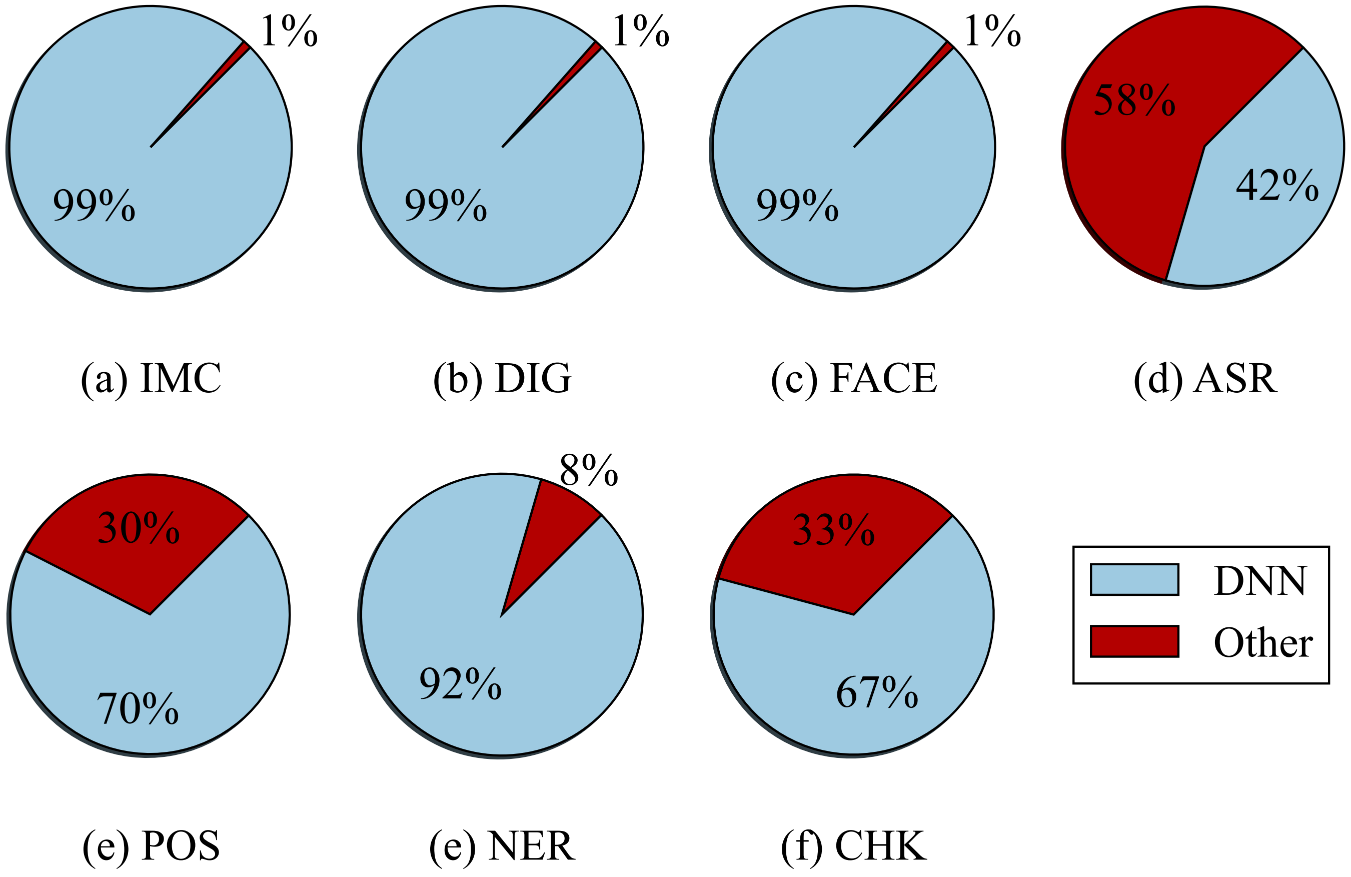
Release includes: inputs, pre-trained models, and modified Caffe

**djinn.clarity-lab.org**

# Identifying Bottlenecks for DNN as a Service

# Identifying Bottlenecks for DjiNN and Tonic

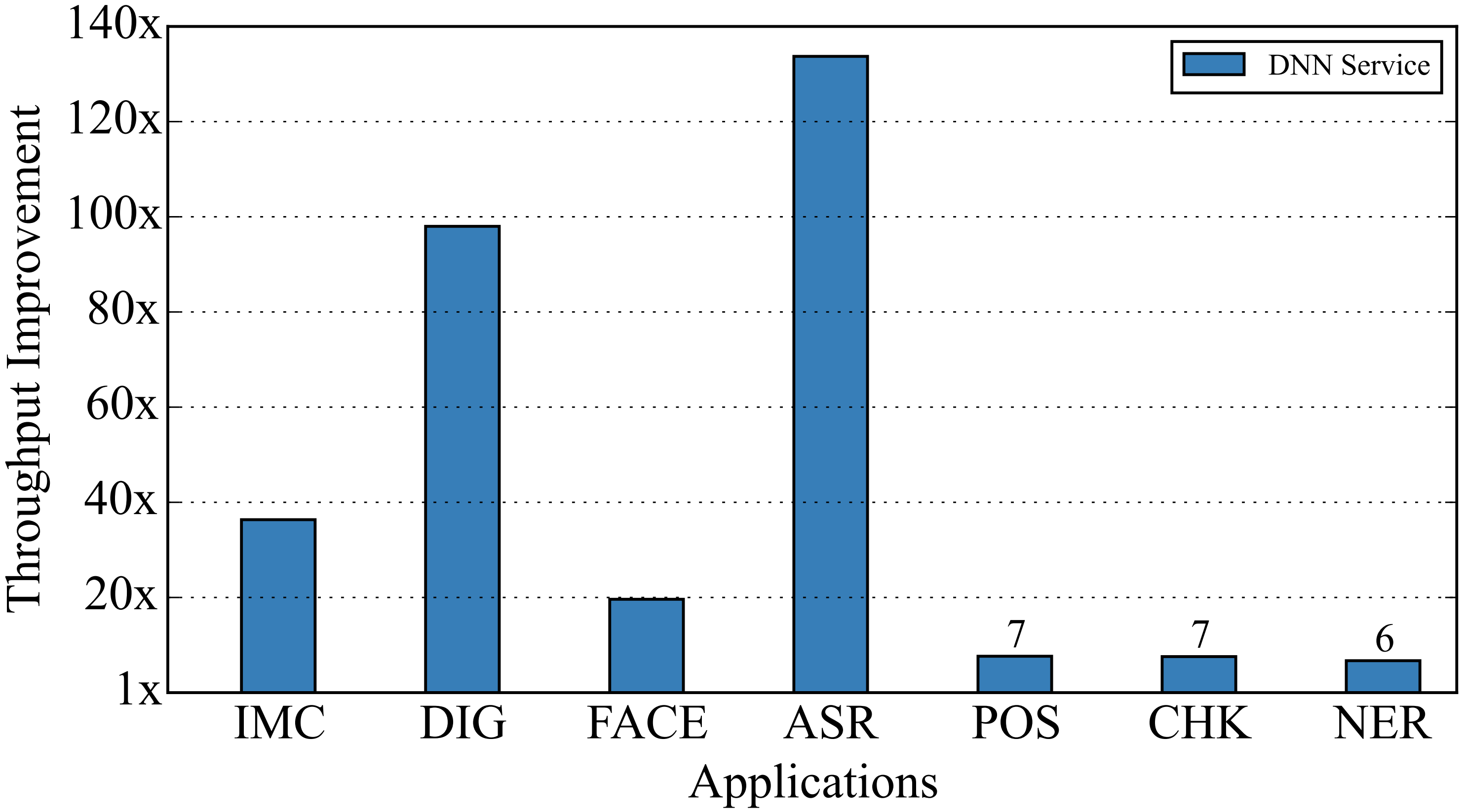
- Software: Caffe (modified)
- CPU: Intel Xeon E5-2620 2.10GHz
  - ATLAS (vectorized)



**DNN: More than 80% of cycles**

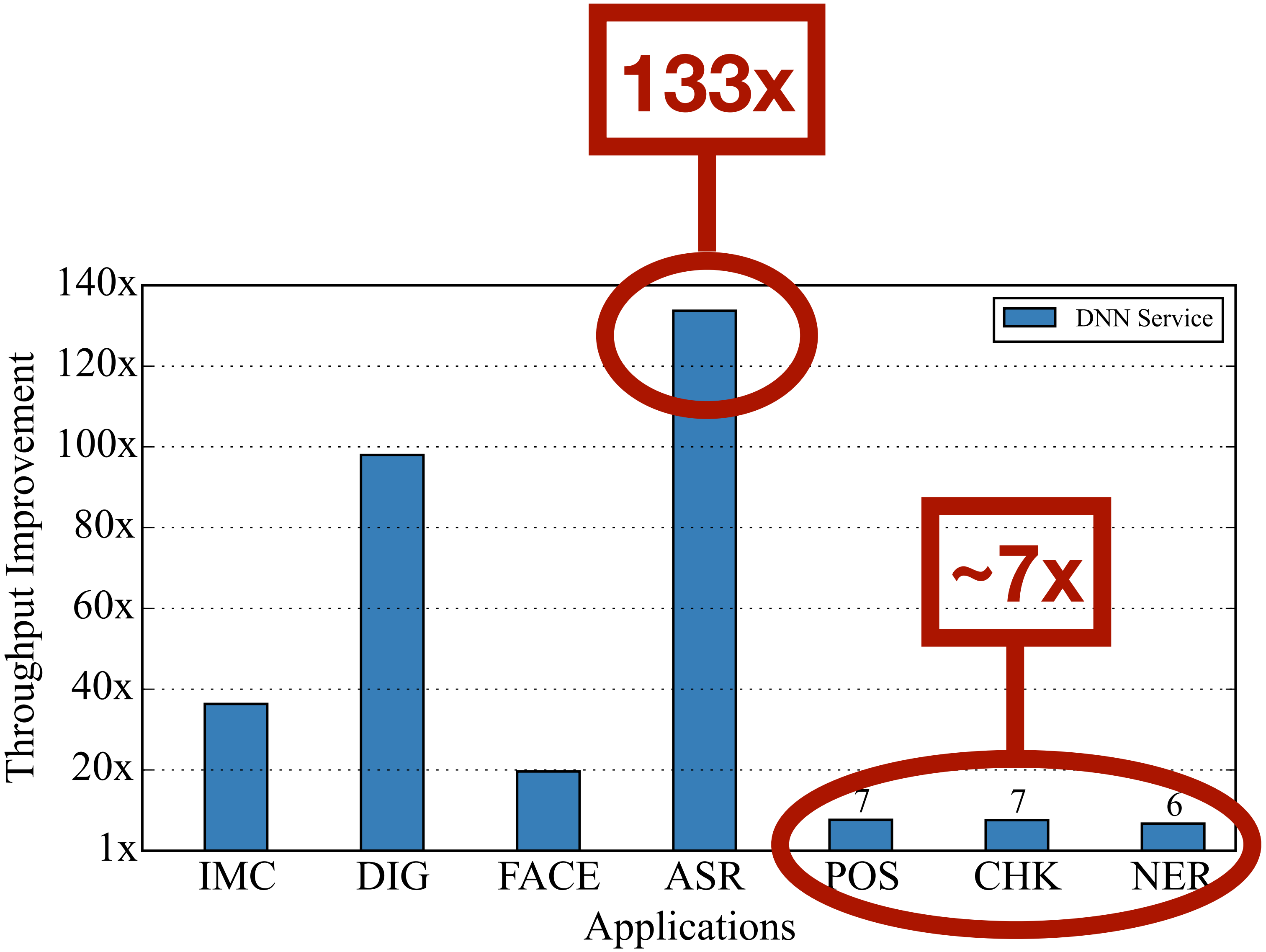
# Identifying Bottlenecks for DjiNN and Tonic

- Software: Caffe (modified)
- CPU: Intel Xeon E5-2620 2.10GHz
  - ATLAS (vectorized)
- GPU: NVIDIA Tesla K40M
  - cuDNN v1 and Caffe

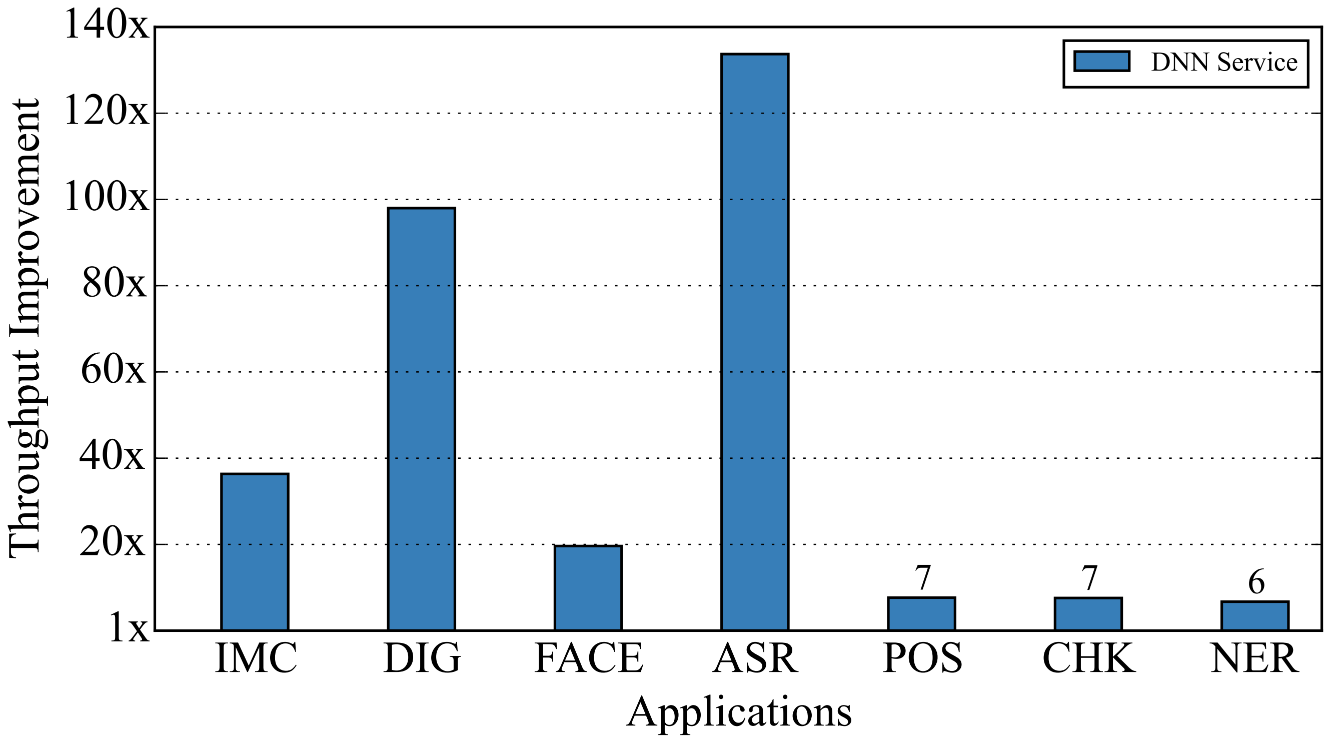


# Identifying Bottlenecks for DjiNN and Tonic

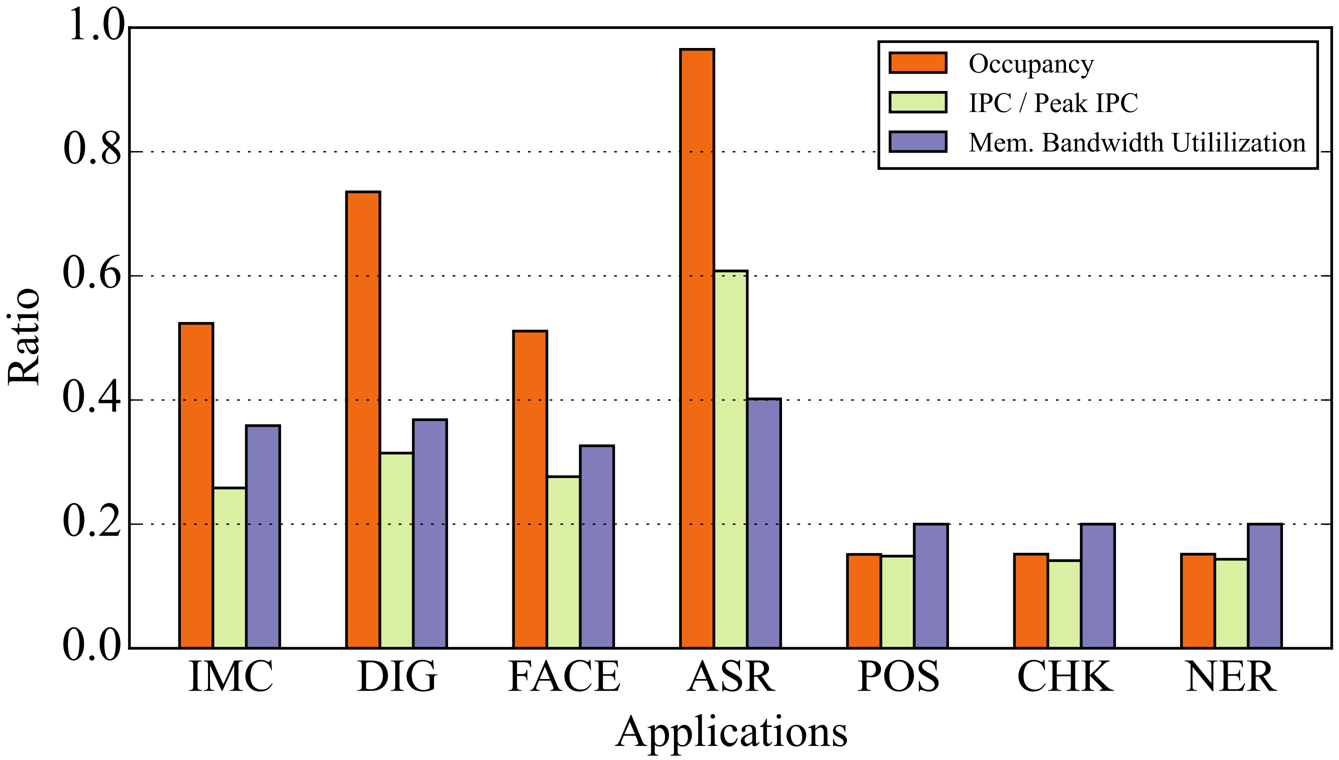
- Software: Caffe (modified)
- CPU: Intel Xeon E5-2620 2.10GHz
  - ATLAS (vectorized)
- GPU: NVIDIA Tesla K40M
  - cuDNN v1 and Caffe



# Identifying Bottlenecks for DjiNN and Tonic

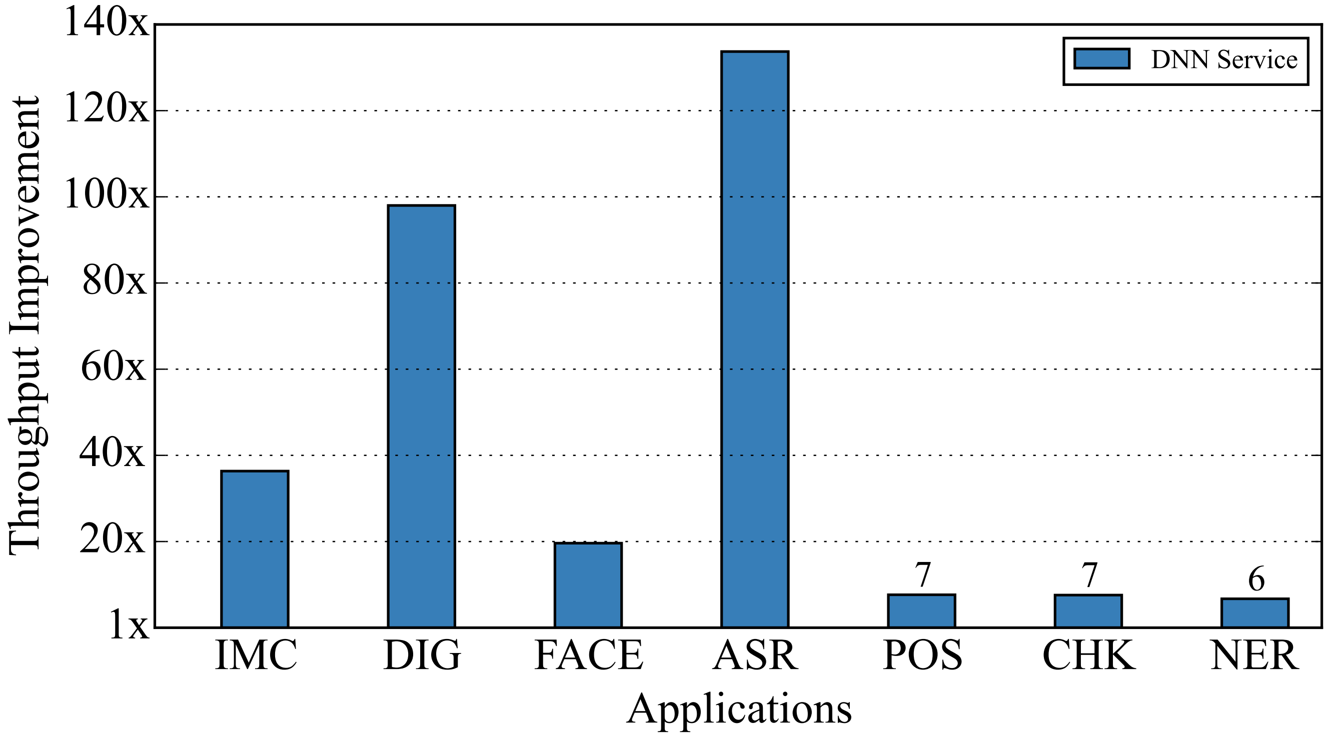


Throughput Improvement



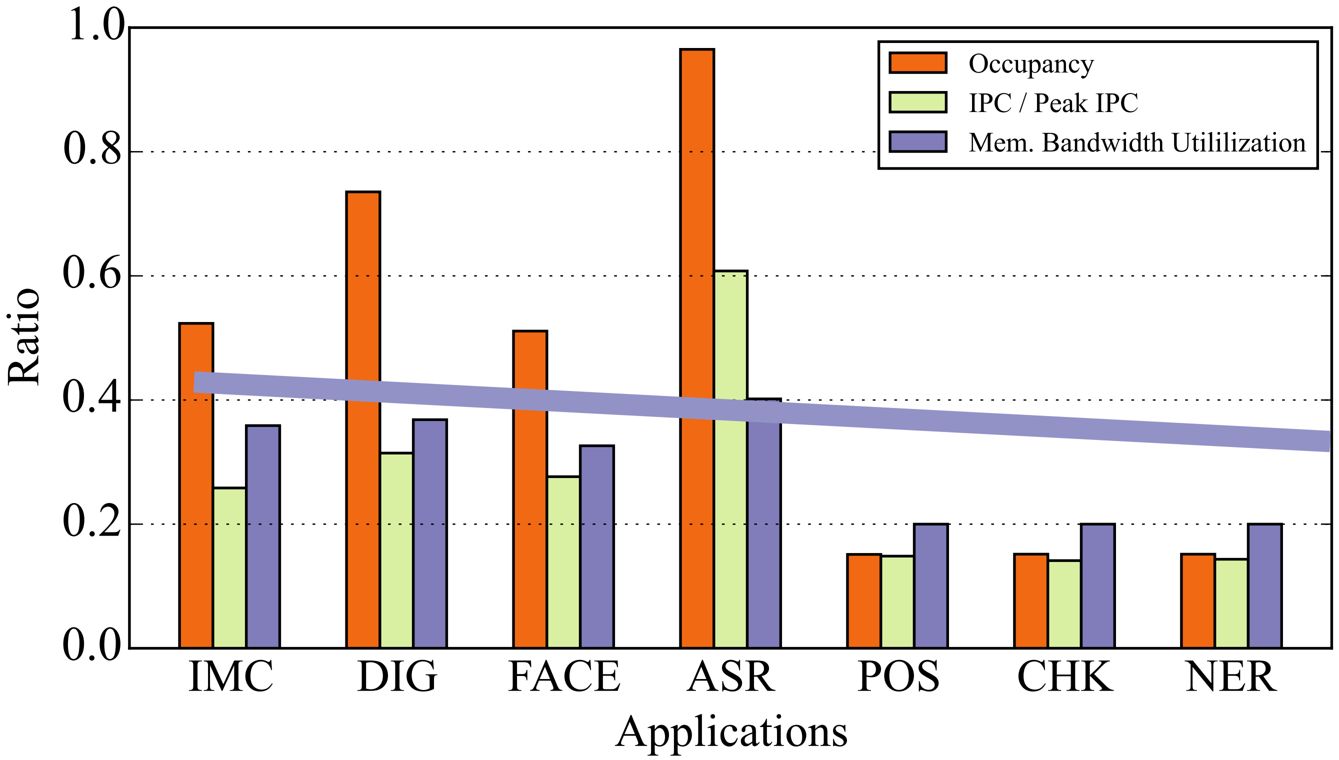
GPU Profiling

# Identifying Bottlenecks for DjiNN and Tonic



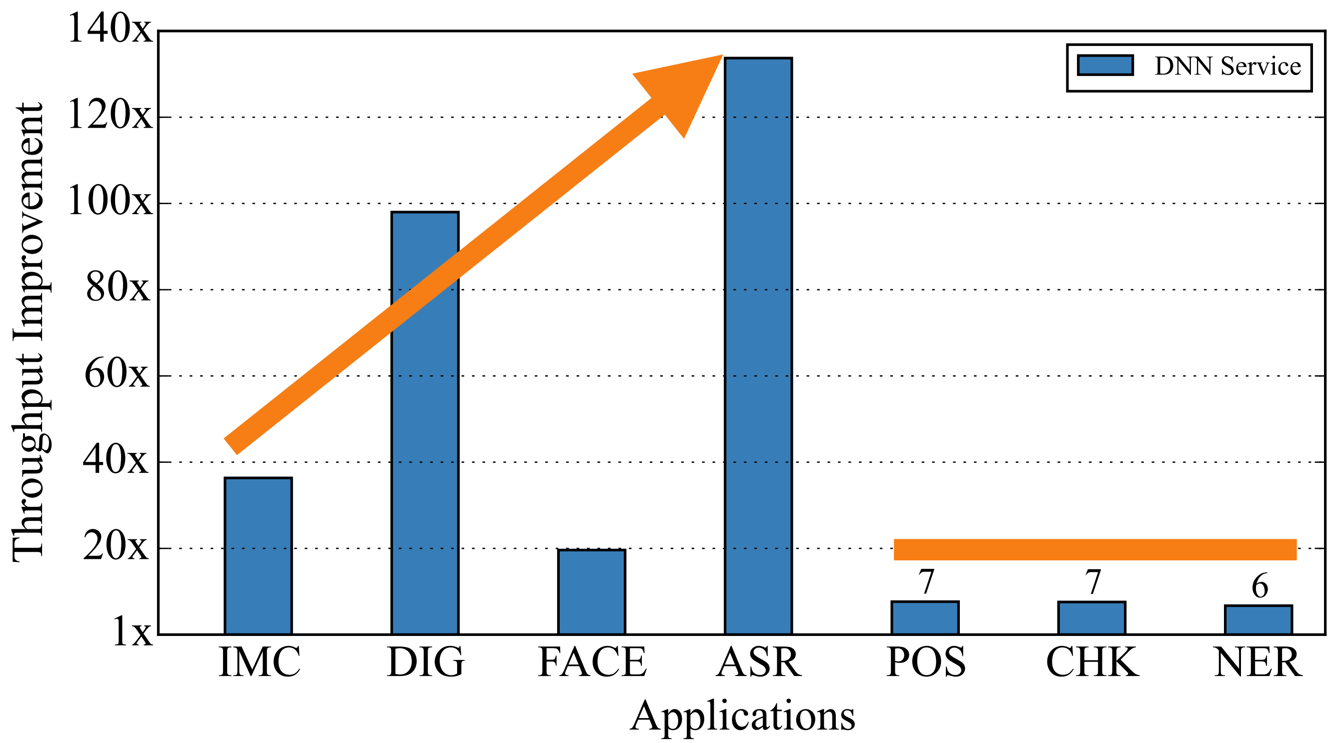
Throughput Improvement

**Low memory bandwidth utilization**

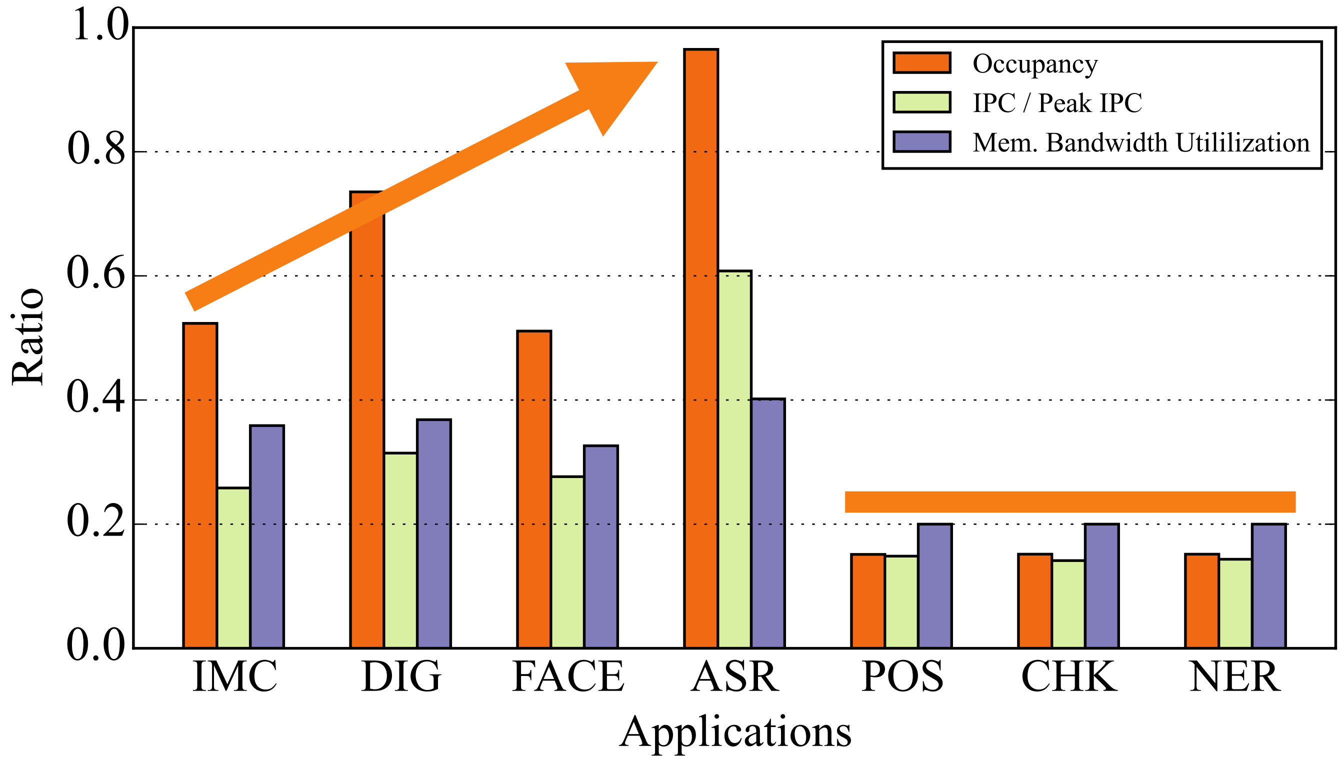


GPU Profiling

# Identifying Bottlenecks for DjiNN and Tonic



Throughput Improvement



GPU Profiling

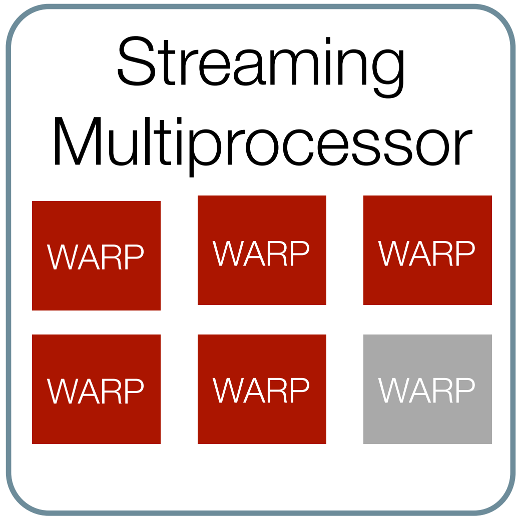
**GPU is not fully utilized**

# Designing a High Throughput System

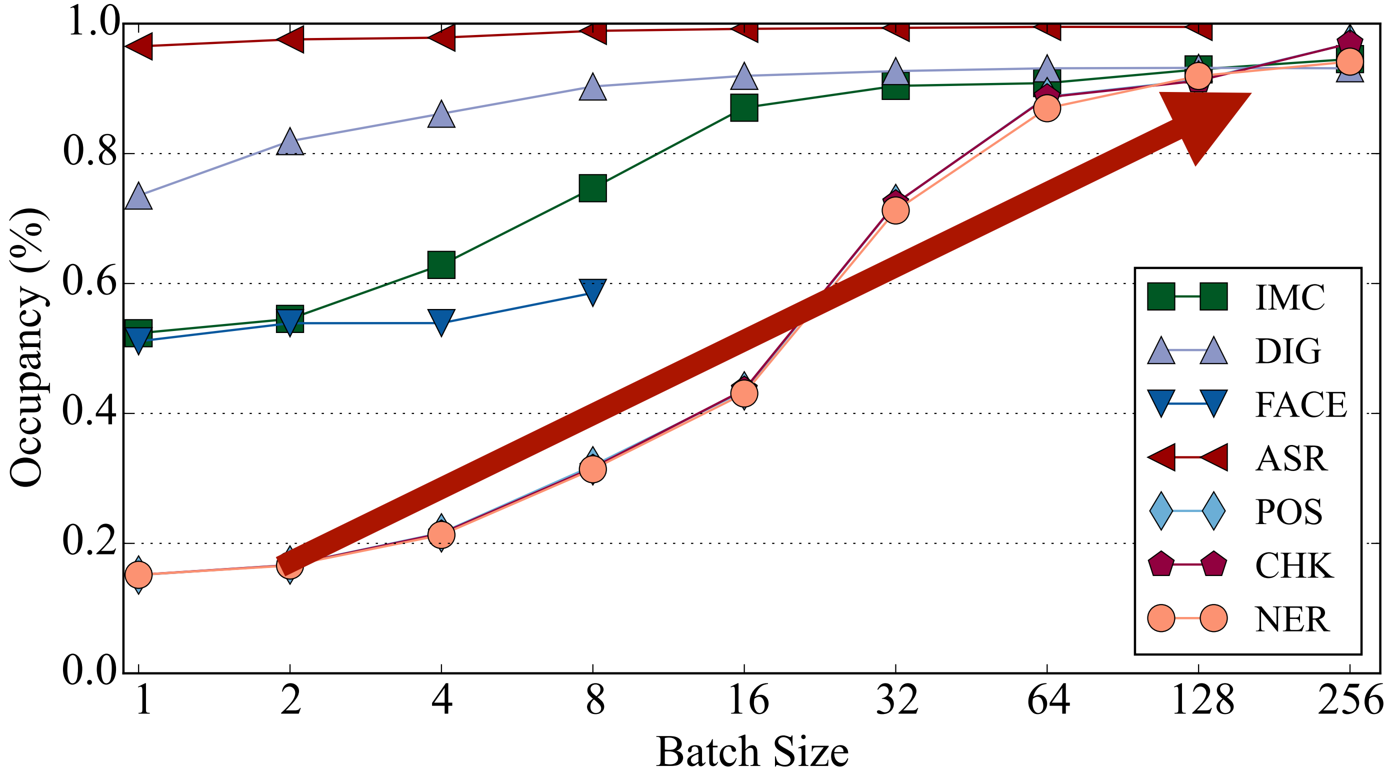
# Designing a High Throughput System — Batching

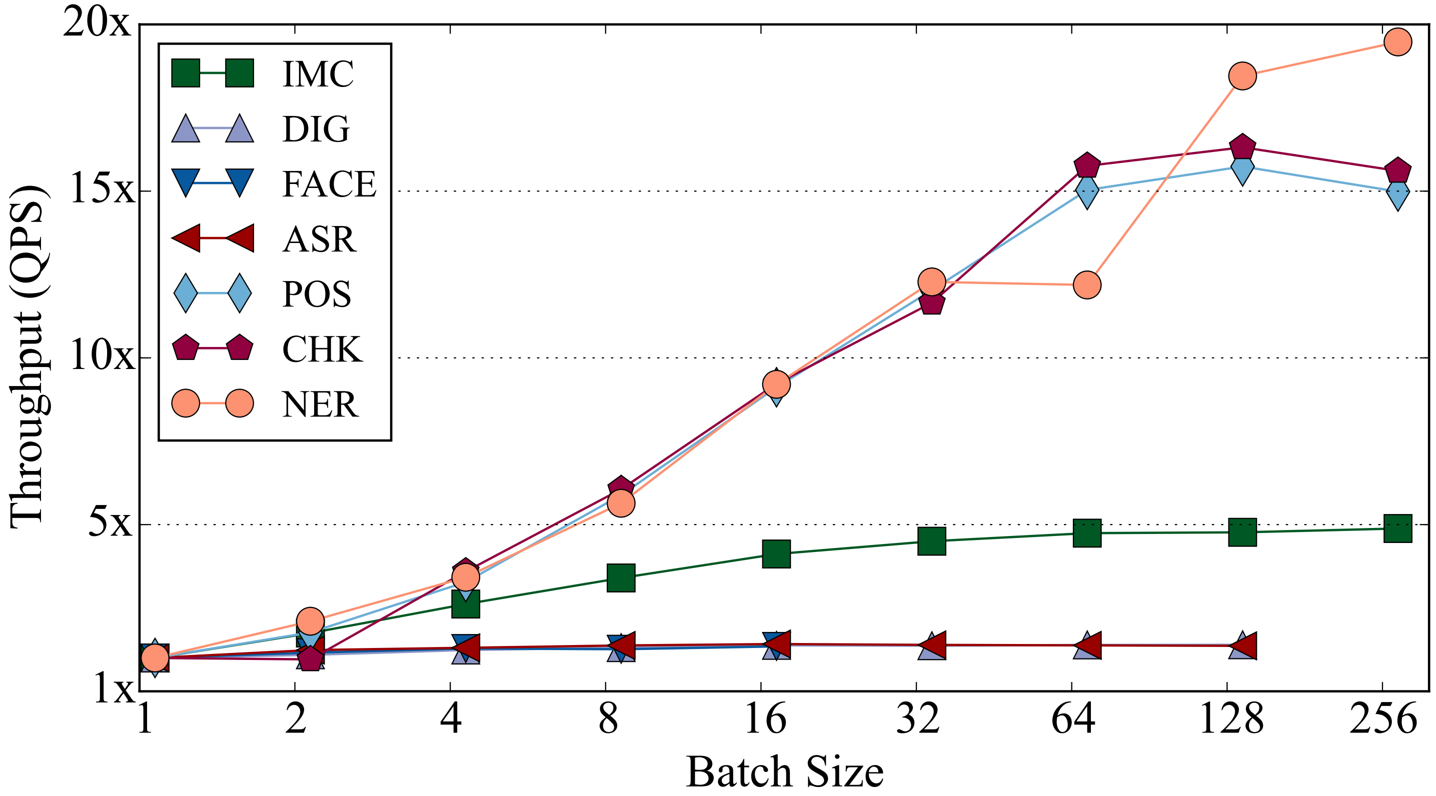
Batch Size:

$$\begin{bmatrix} x_0 & x_{01} & x_{02} & x_{03} & x_{04} & x_{05} \\ x_1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_d & x_d & x_d & x_d & x_d & x_d \end{bmatrix}$$


Streaming Multiprocessor

✔ High Occupancy

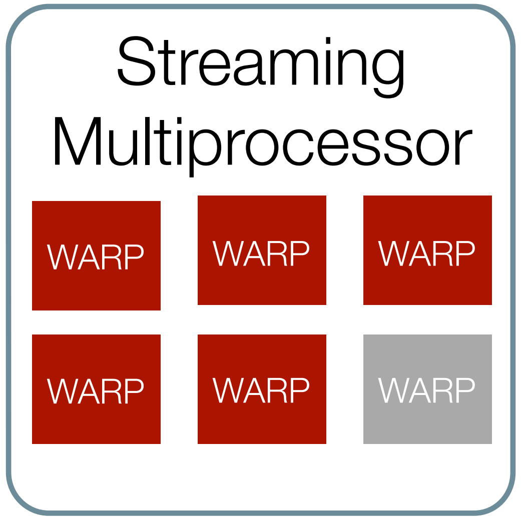




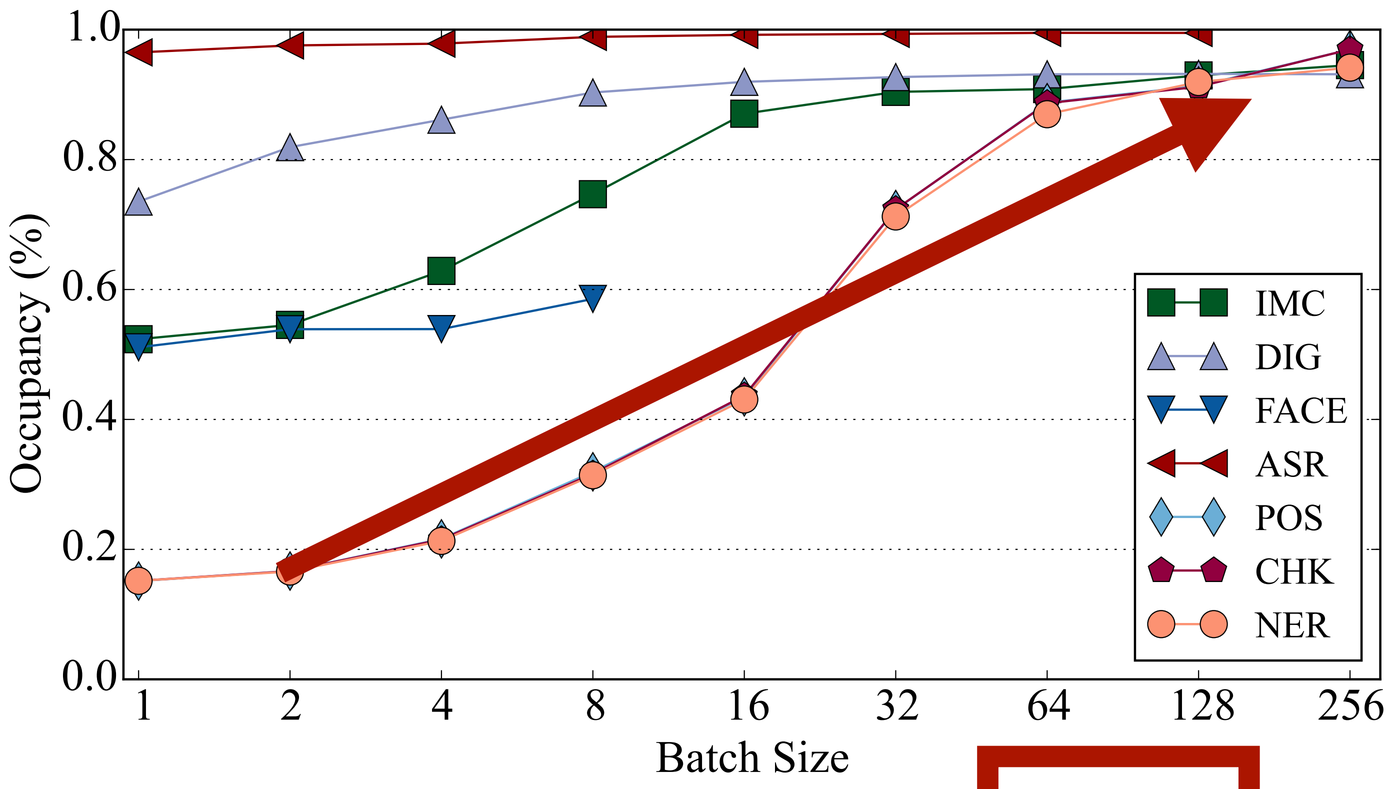
# Designing a High Throughput System — Batching

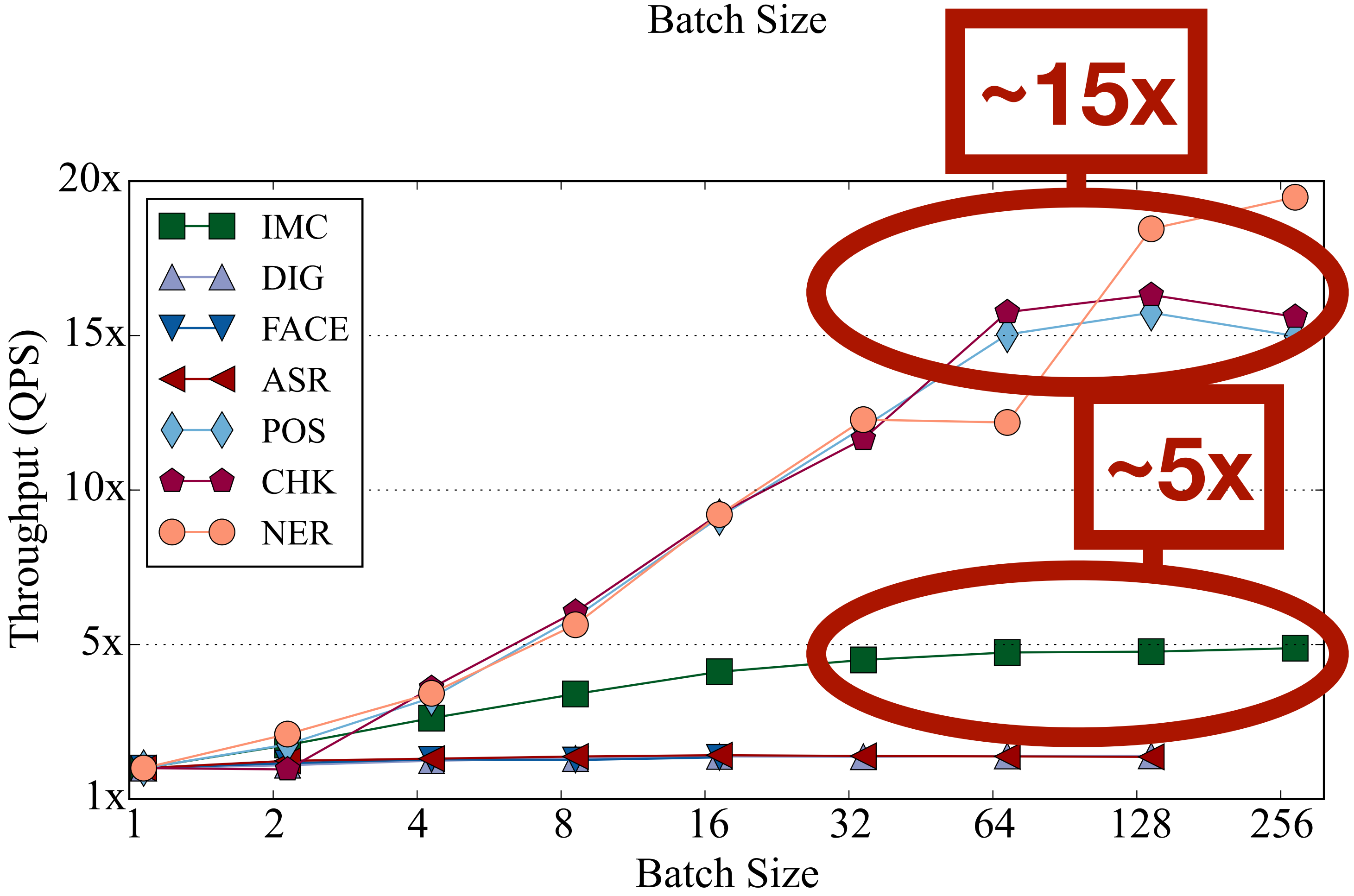
Batch Size:

$$\begin{bmatrix} X_0 & X_{01} & X_{02} & X_{03} & X_{04} & X_{05} \\ X_1 & X_{11} & X_{12} & X_{13} & X_{14} & X_{15} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_d & X_d & X_d & X_d & X_d & X_d \end{bmatrix}$$



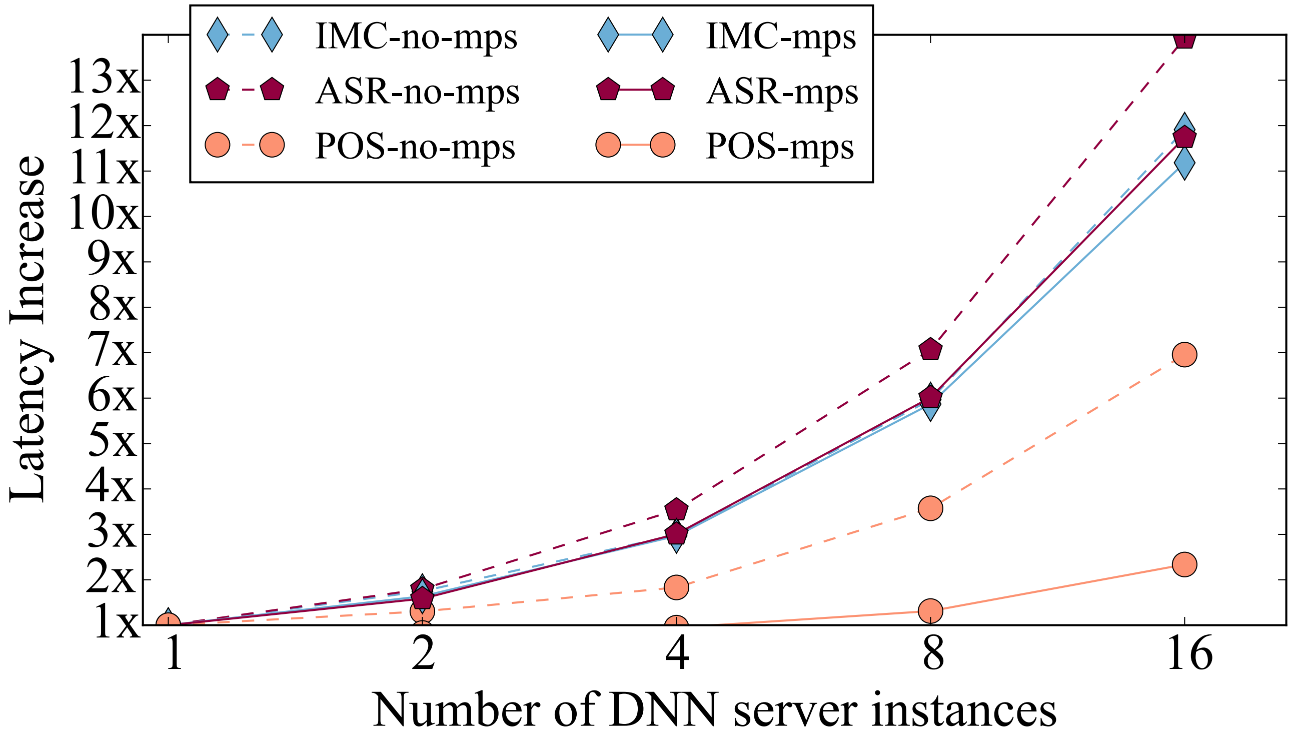
✔ High Occupancy

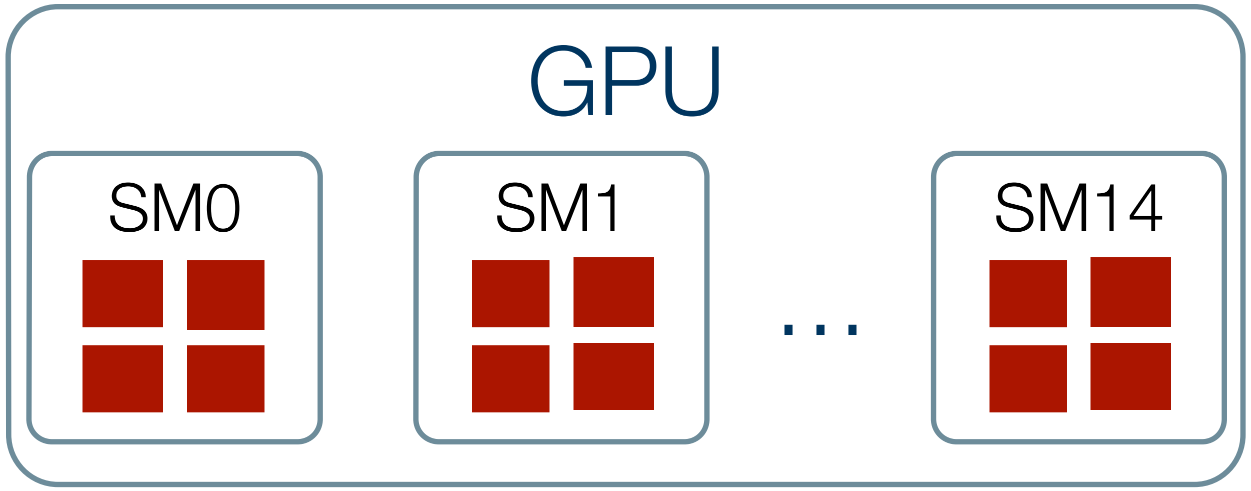


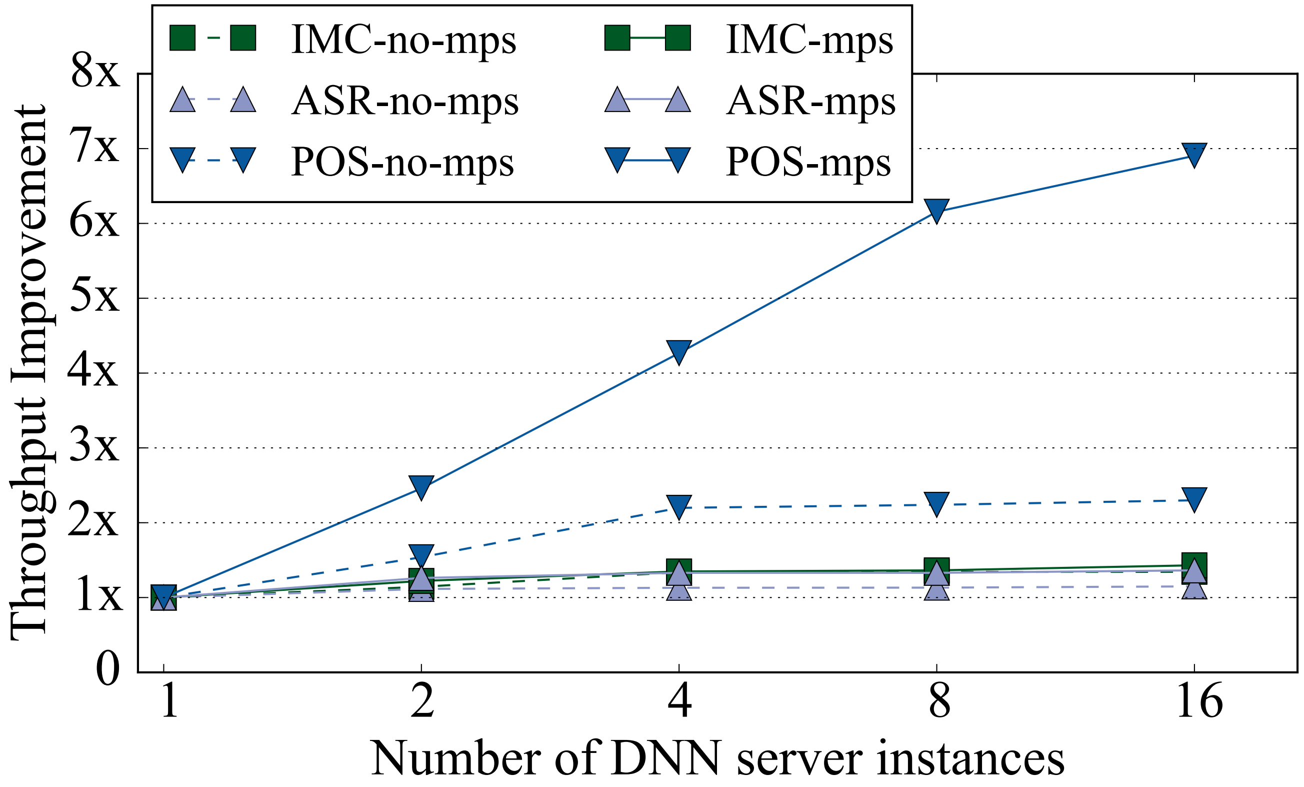


# Designing a High Throughput System — Concurrent Execution

- Launch concurrent DNN services on the GPU
- Leverage NVIDIA Multi-Process Service (MPS) [1]



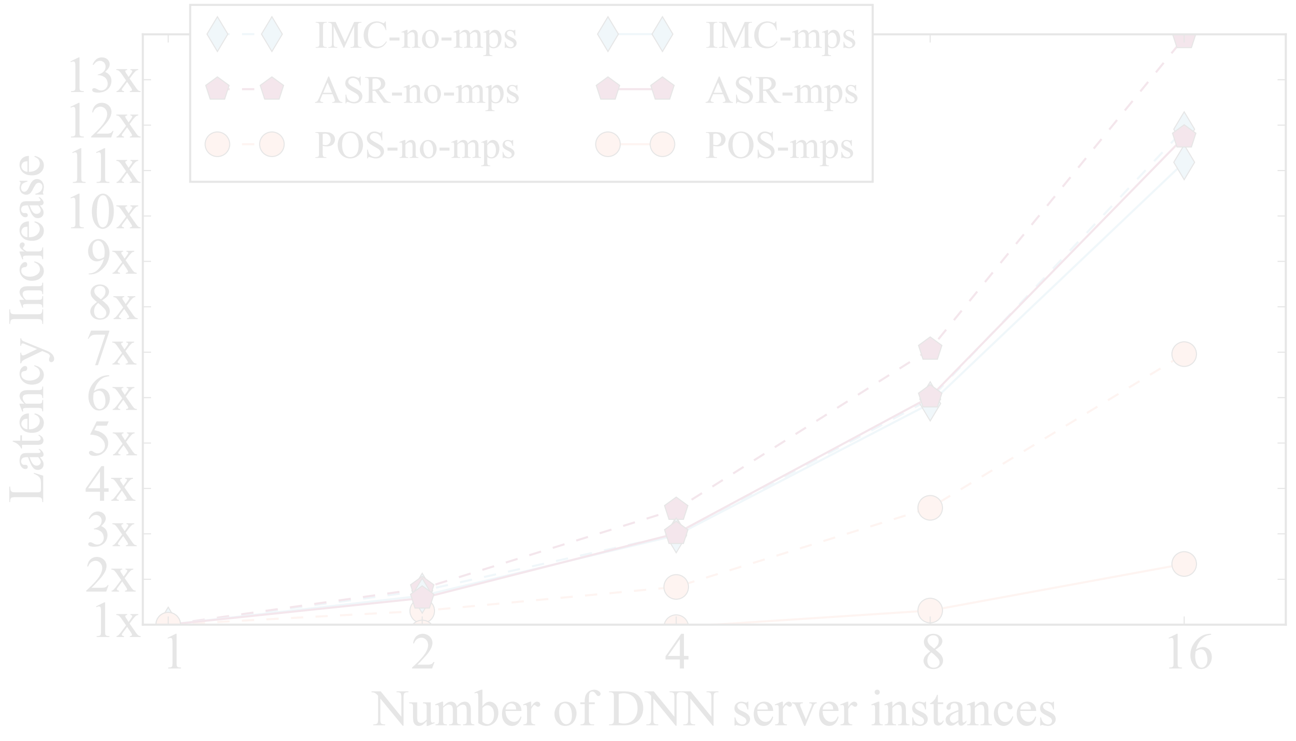




[1] “Multi-Process Service” https://docs.nvidia.com/deploy/pdf/CUDA_ Multi_Process_Service_Overview.pdf

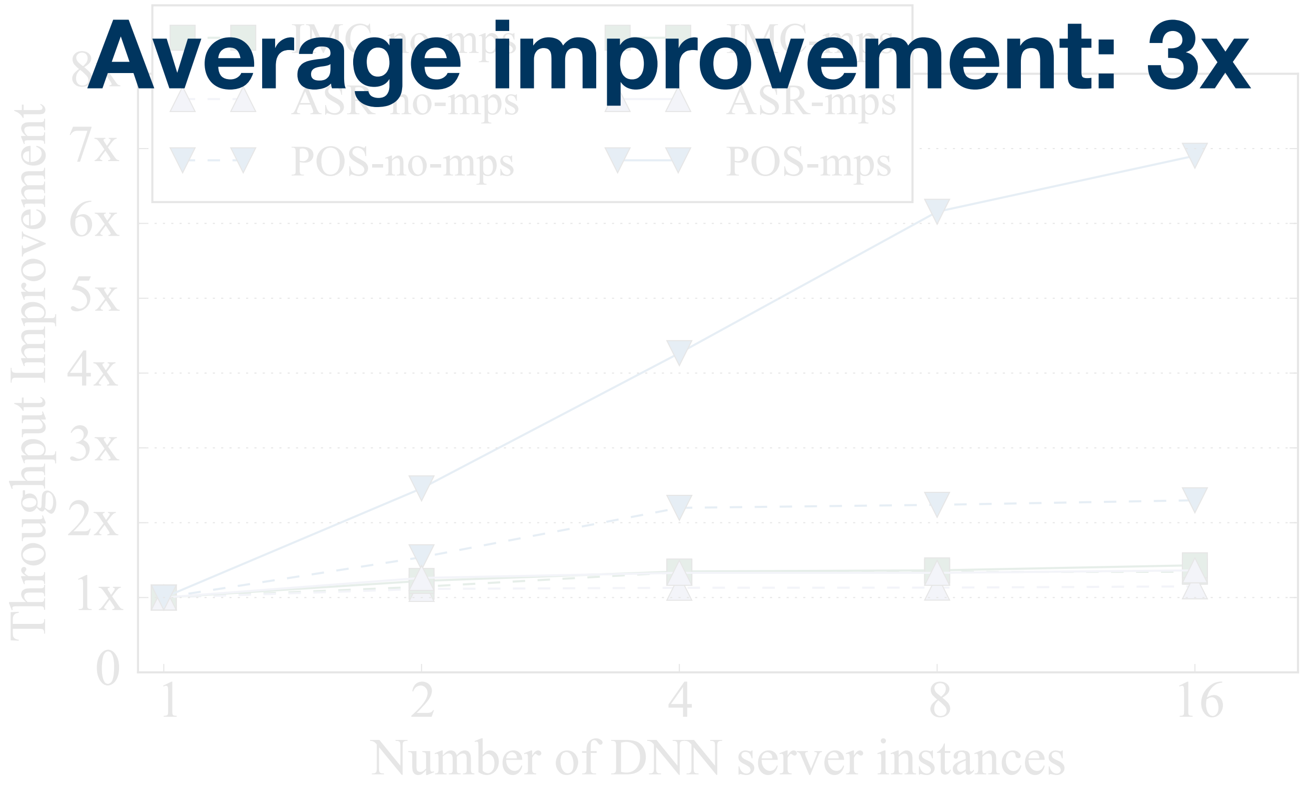
# Designing a High Throughput System — Concurrent Execution

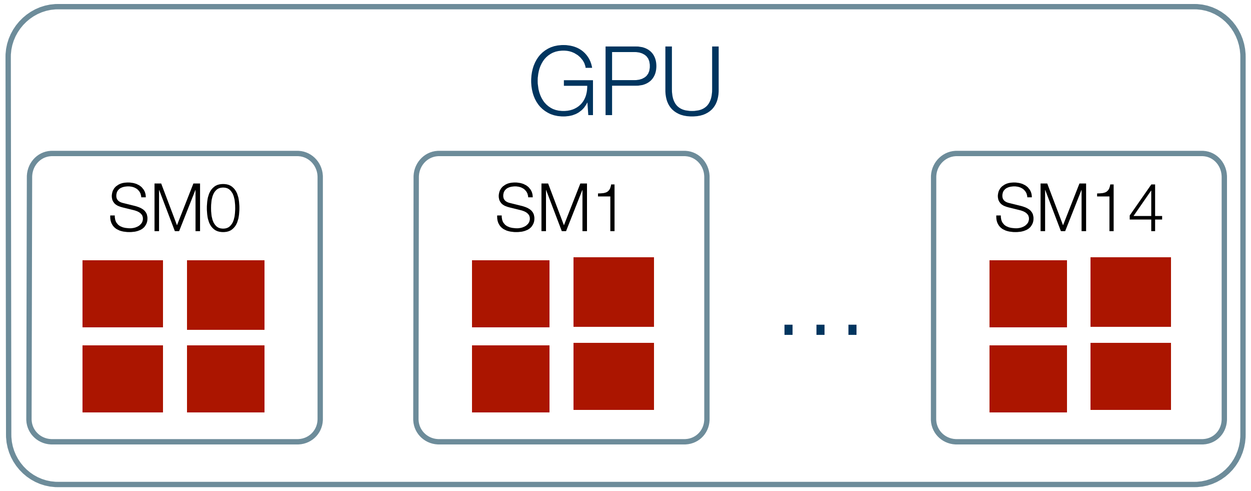
- Launch concurrent DNN services on the GPU
- Leverage NVIDIA Multi-Process Service (MPS) [1]



**4 DNN concurrent instances**

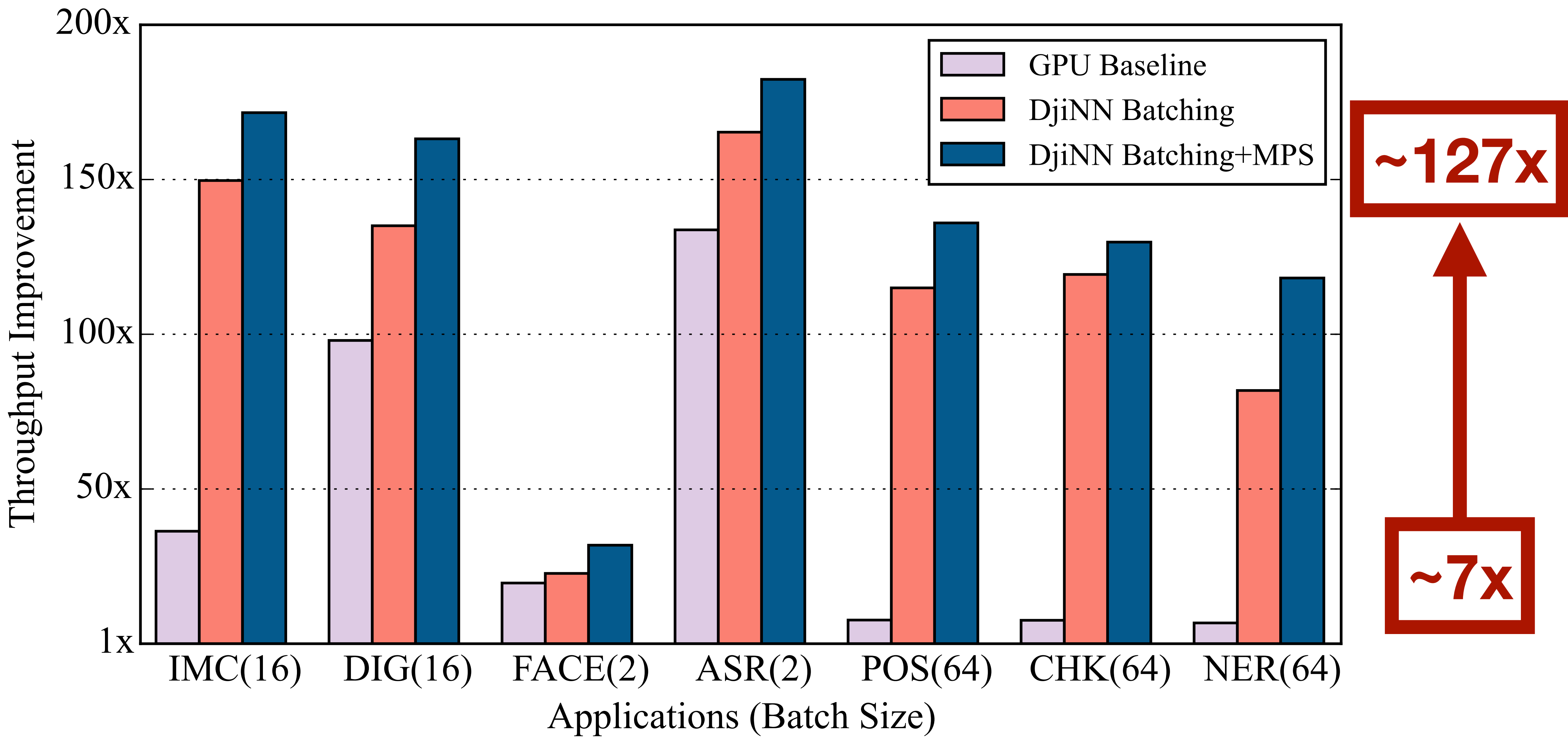
**Average improvement: 3x**



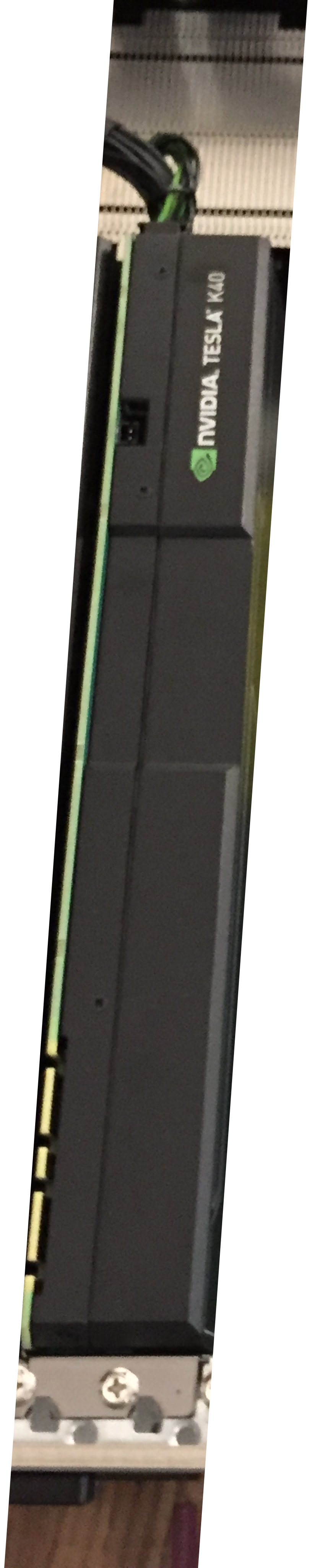


[1] “Multi-Process Service” https://docs.nvidia.com/deploy/pdf/CUDA_ Multi_Process_Service_Overview.pdf

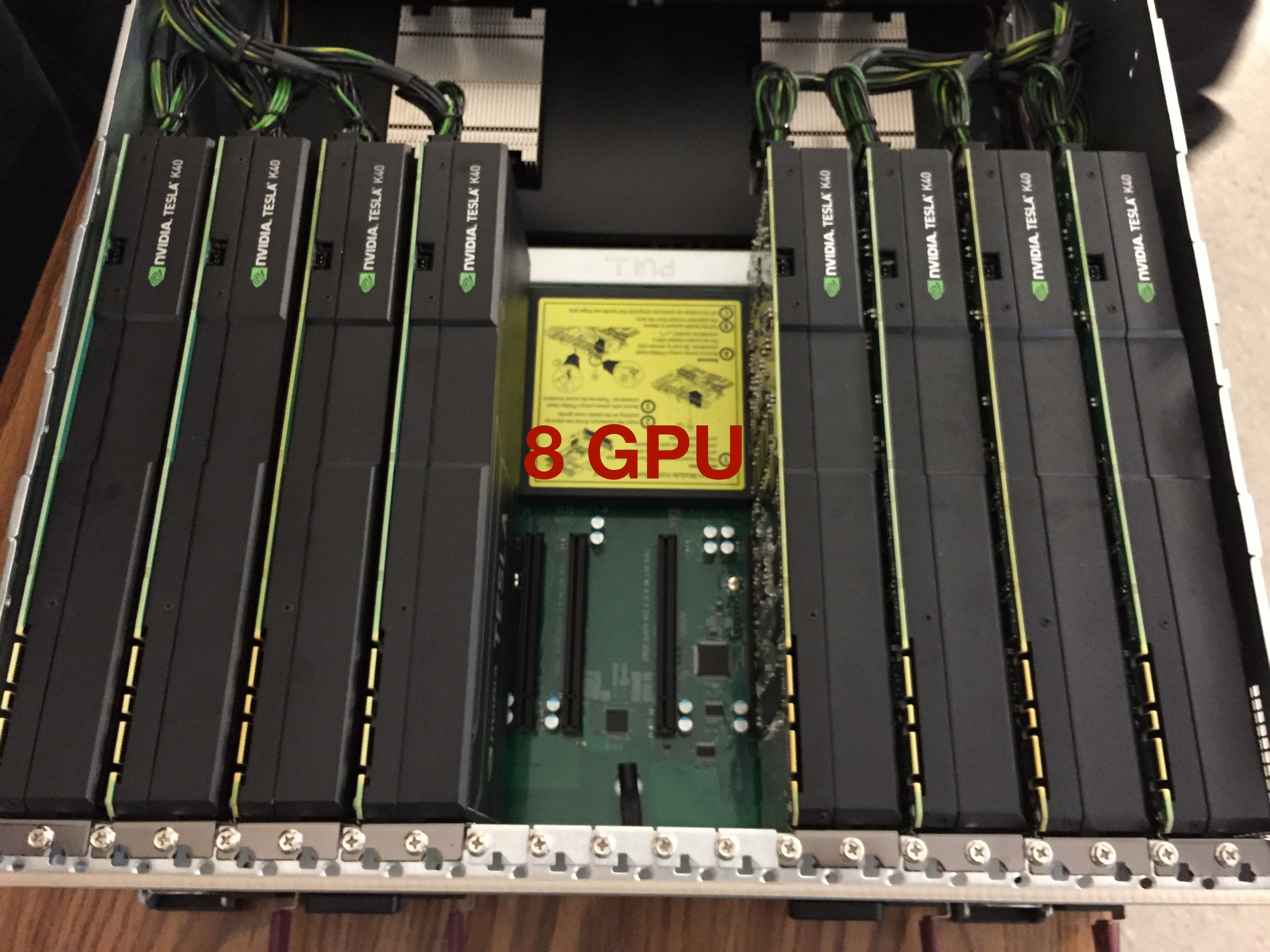
# Designing a High Throughput System



**Average throughput improvement: 133x**

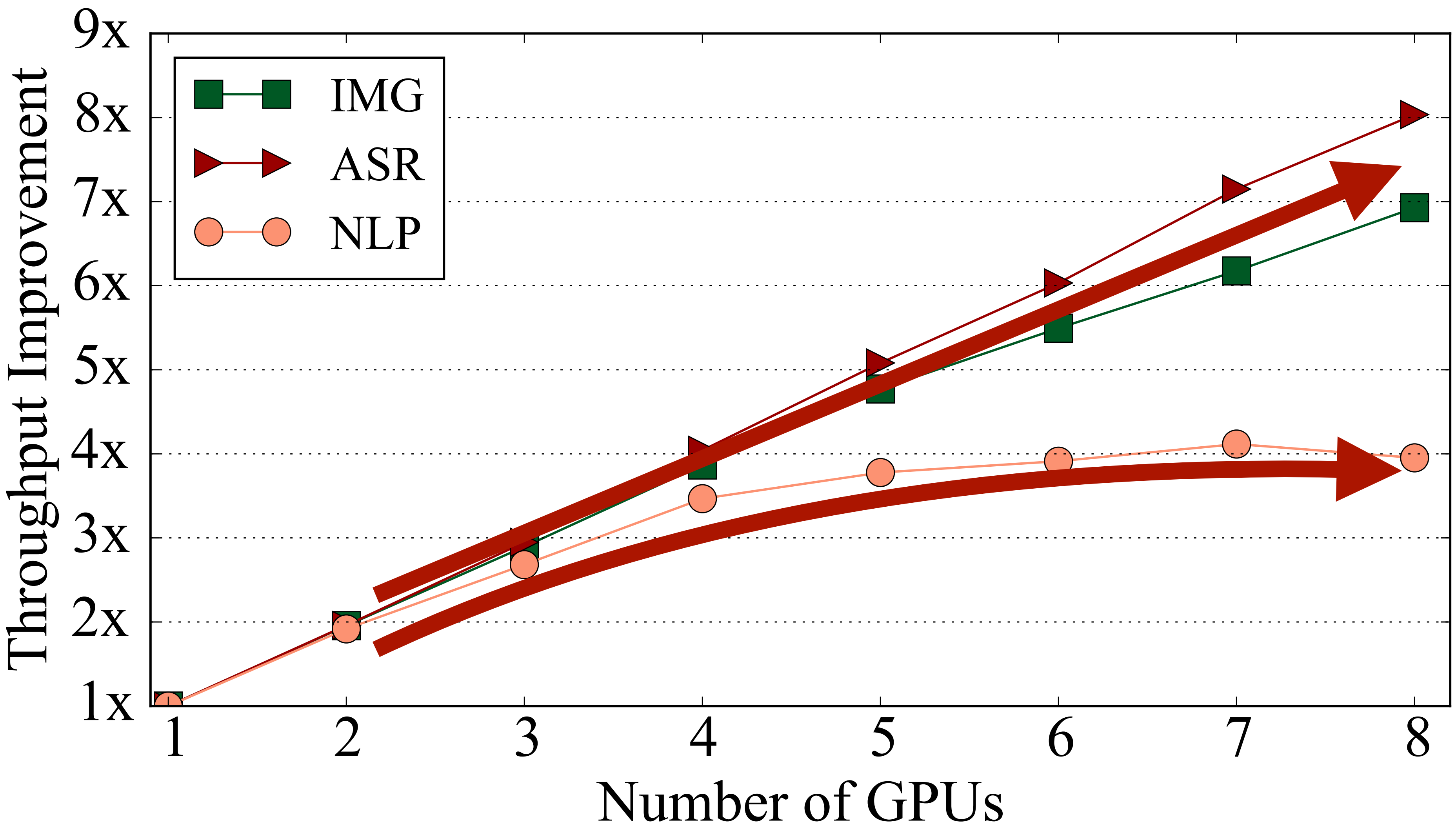


1 GPU


NVIDIA TESLA K40
NVIDIA TESLA K40
NVIDIA TESLA K40
NVIDIA TESLA K40
NVIDIA TESLA K40
NVIDIA TESLA K40
NVIDIA TESLA K40
NVIDIA TESLA K40
PULL

8 GPU

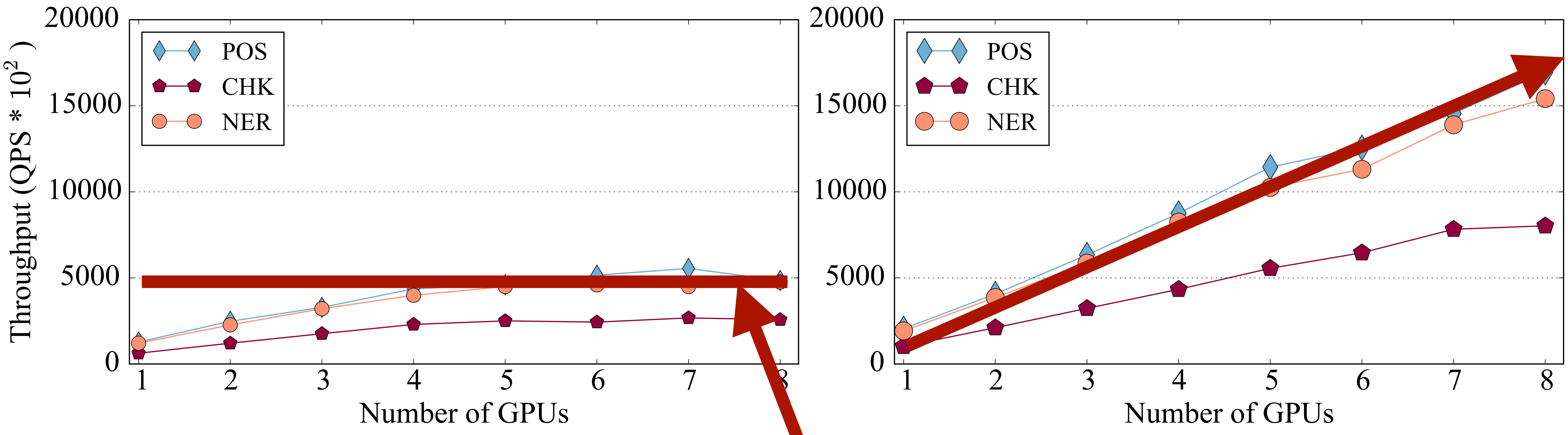
# GPU Scaling



**Average throughput improvement using optimizations: 771x**

# Bandwidth Requirements for Peak Throughput

Experimental setup: eliminate any data transfer to GPU



**Bandwidth to GPUs saturated**

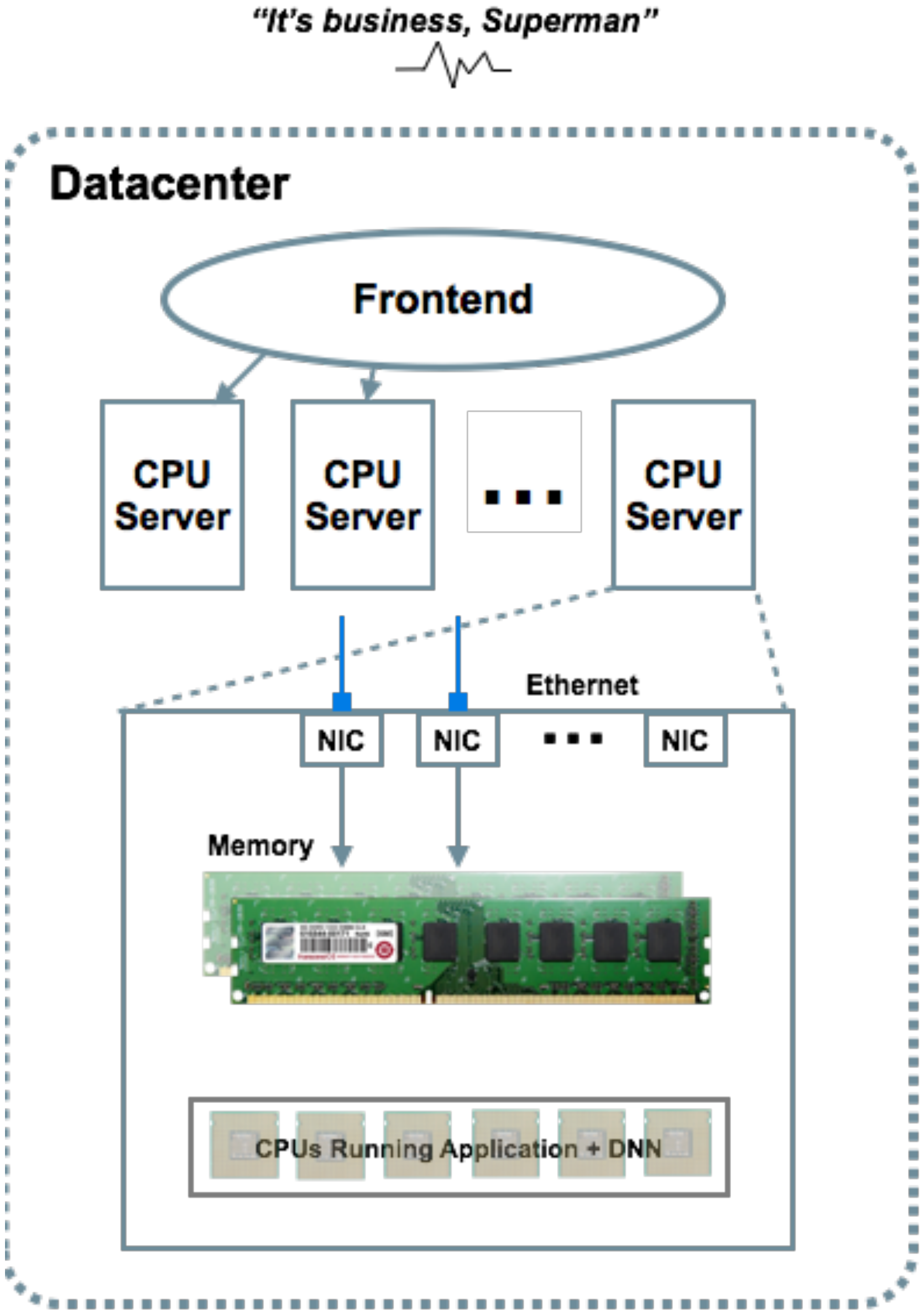
**NLP requires more bandwidth to GPUs**

# Single Server Key Insights

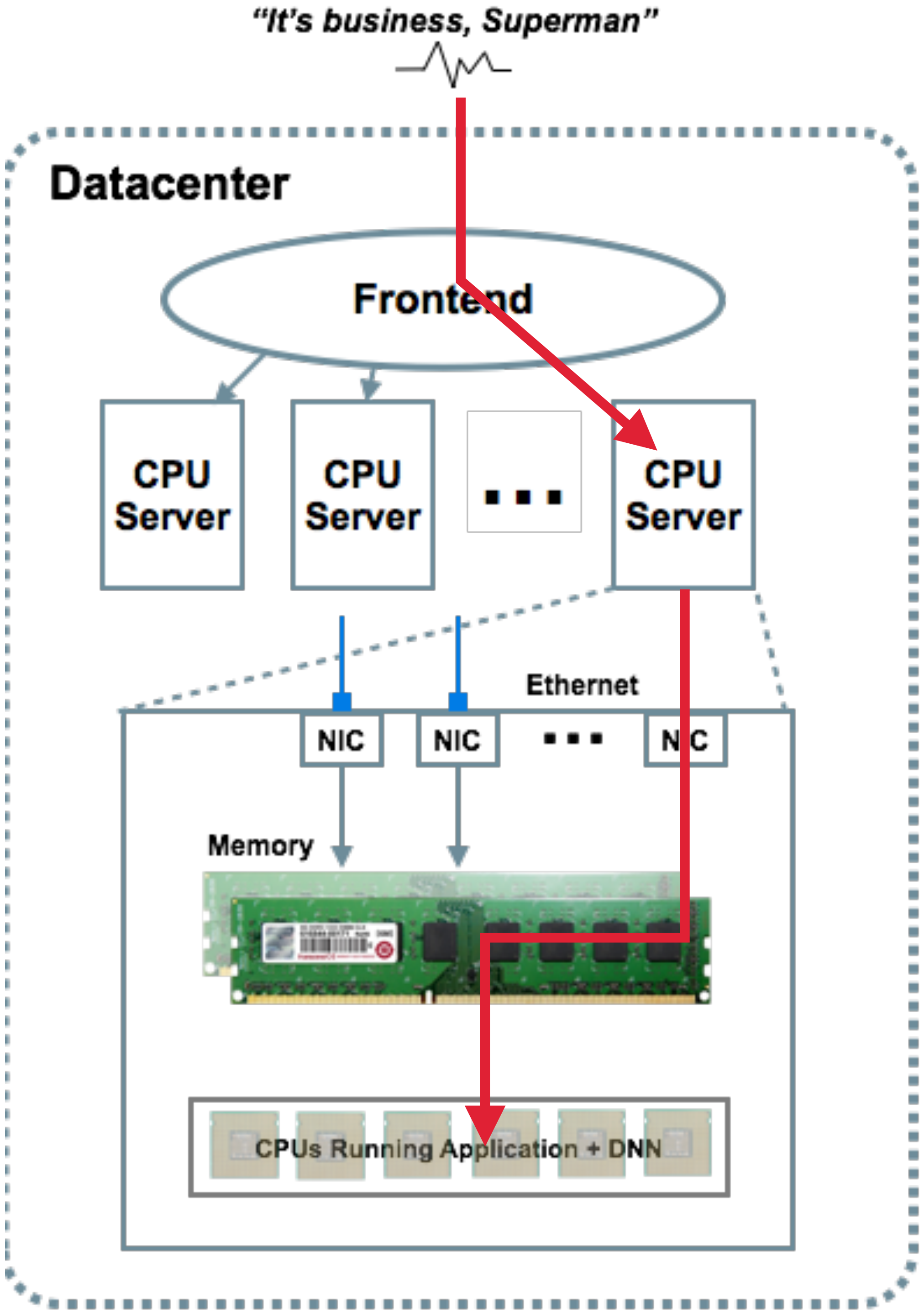
- DNNs do not benefit equally from optimizations
  - High communication Natural Language Processing (NLP) tasks require far more bandwidth
- Optimizing compute platforms depends on DNN’s computation and communication characteristics

# Future Warehouse Scale Computer (WSC) Designs

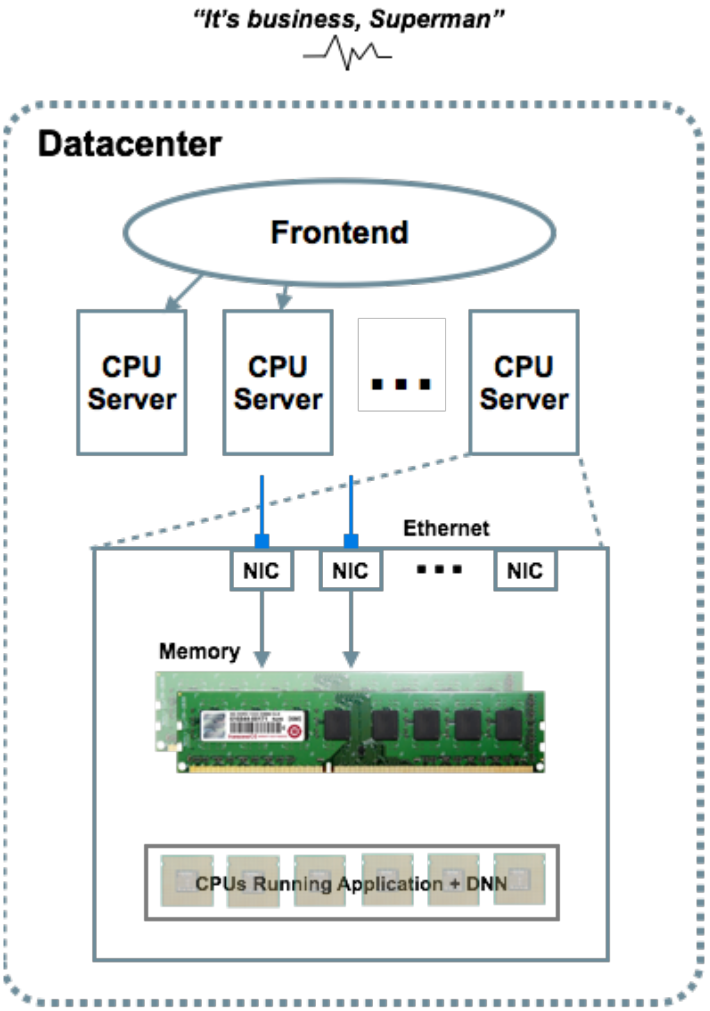
# Future WSC Designs — Server Designs



# Future WSC Designs — Server Designs



# Future WSC Designs — Server Designs

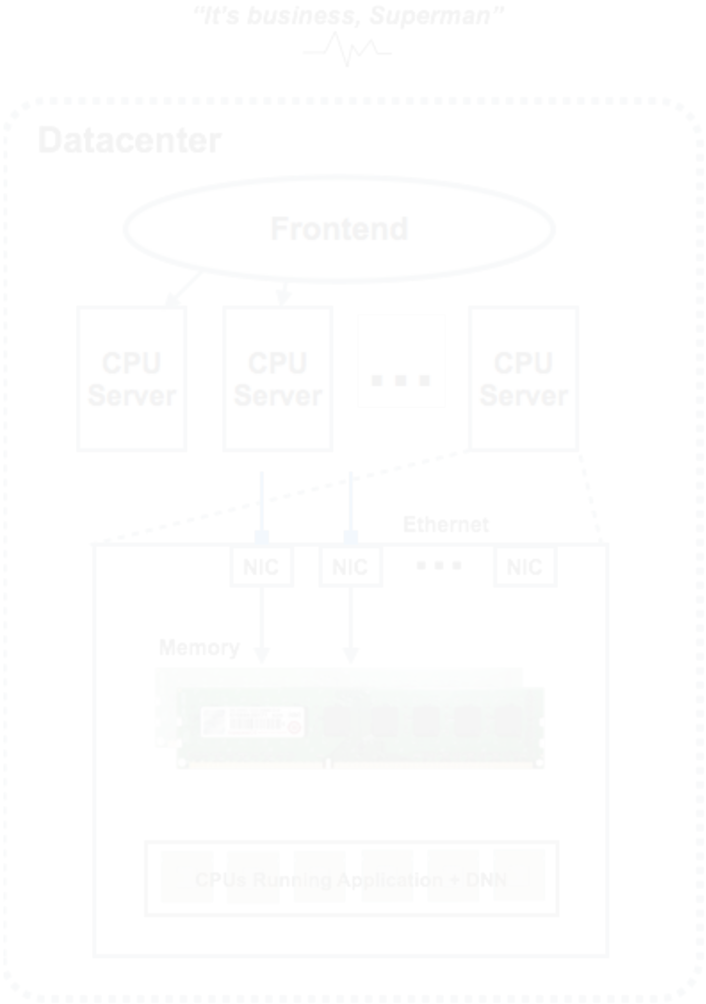


CPU-only:

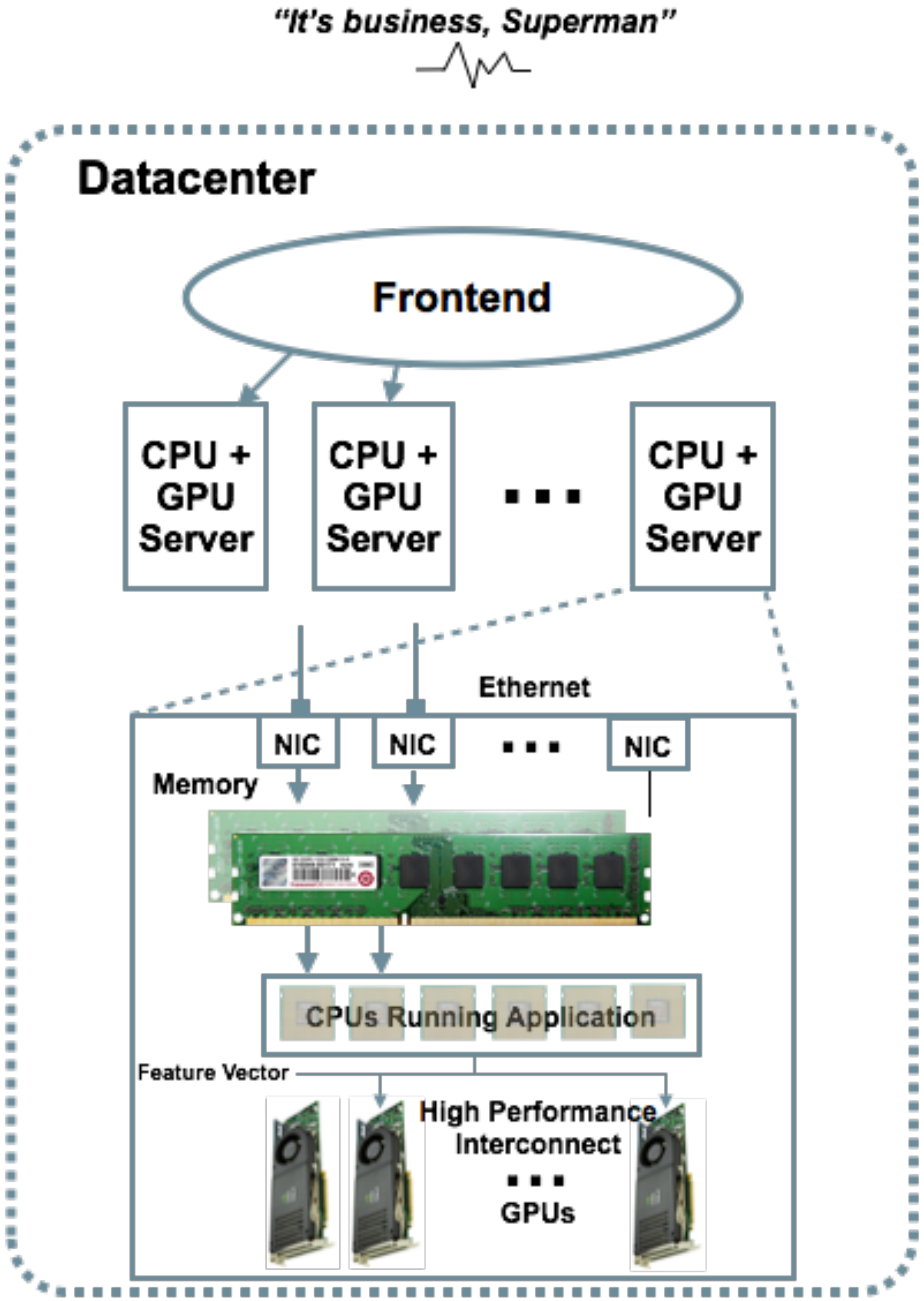
+no extra HW

-low throughput

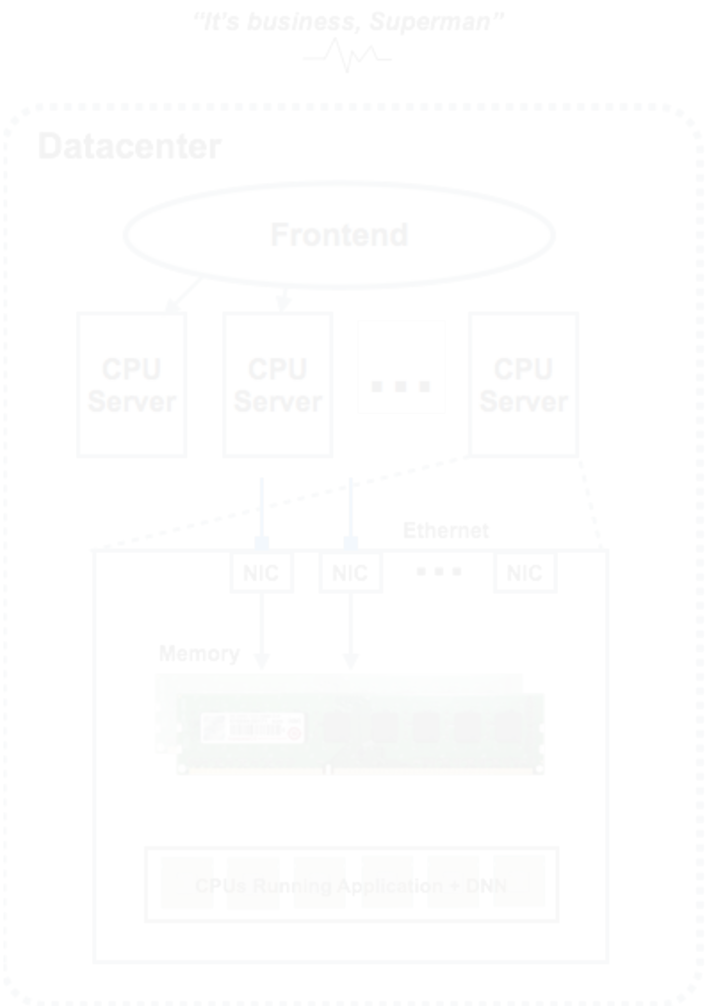
# Future WSC Designs — Server Designs



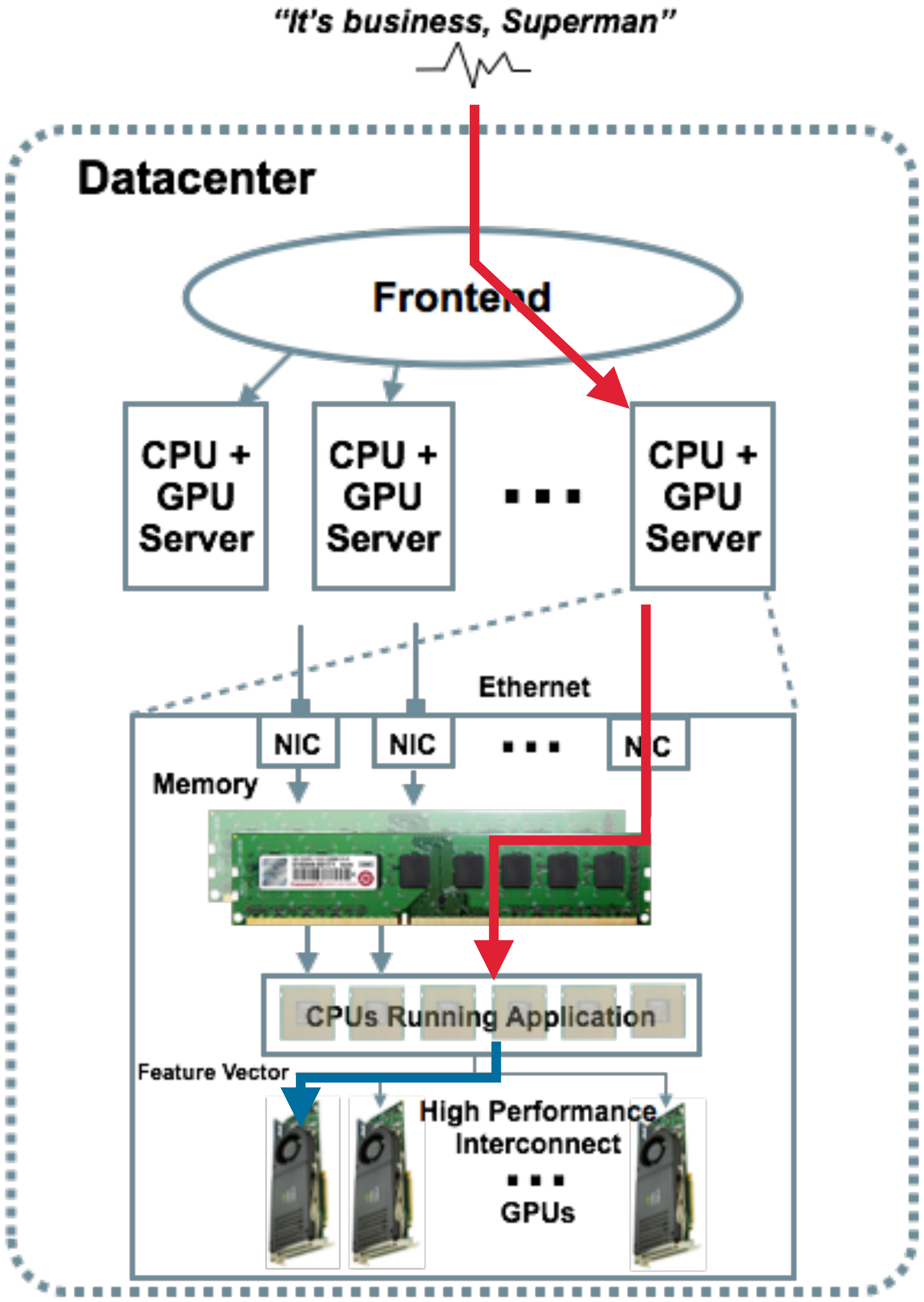
CPU-only:
+no extra HW
-low throughput



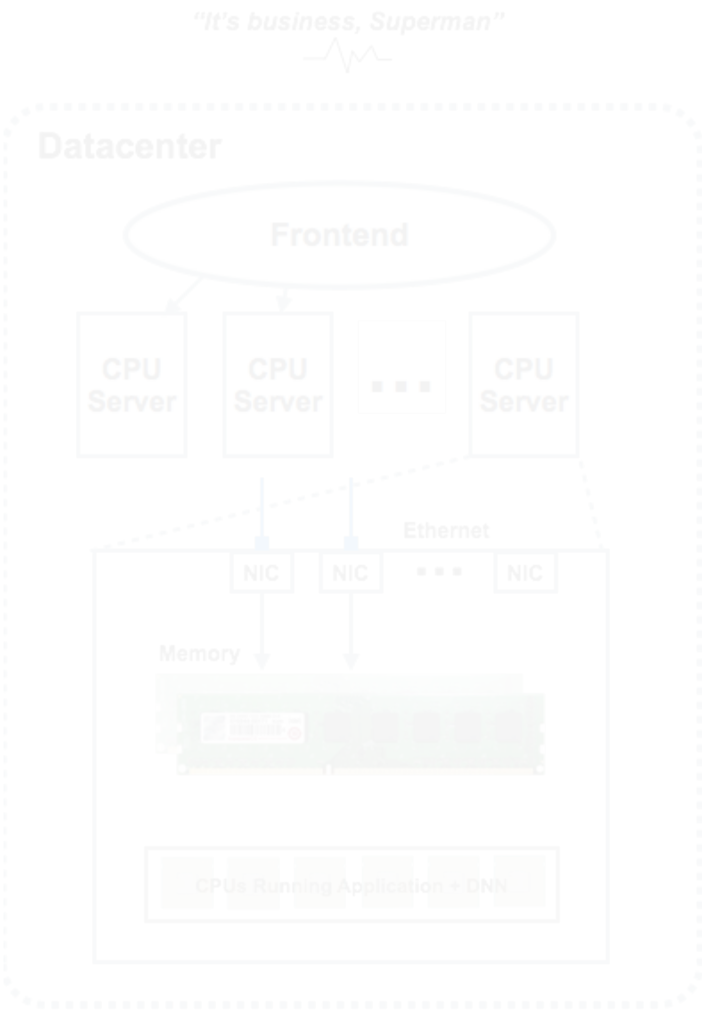
# Future WSC Designs — Server Designs



CPU-only:
+no extra HW
-low throughput



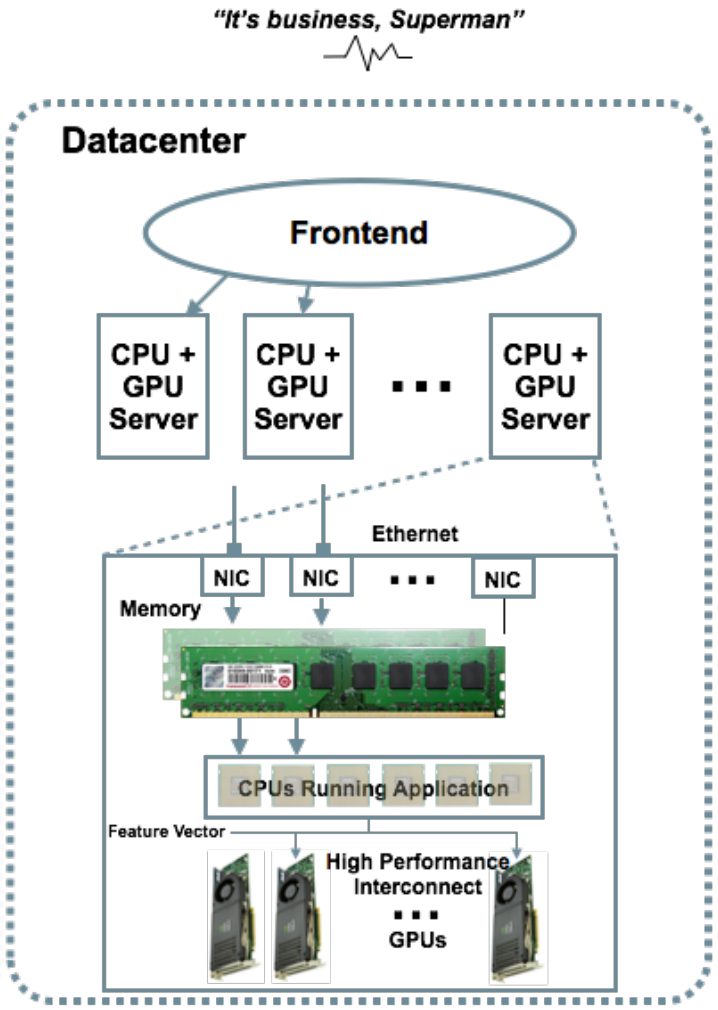
# Future WSC Designs — Server Designs



CPU-only:

+no extra HW

-low throughput

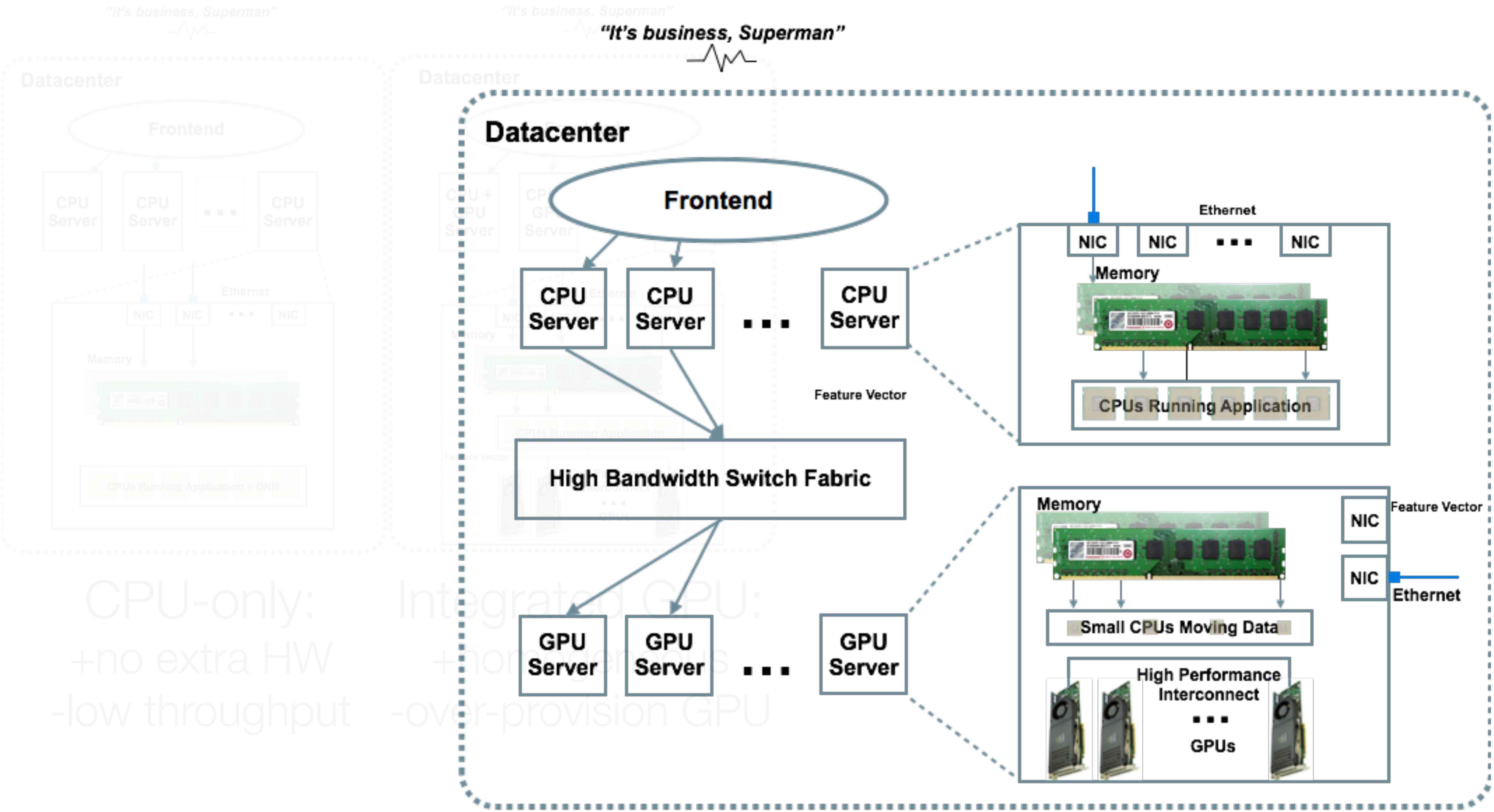


Integrated GPU:

+homogeneous

-over-provision GPU

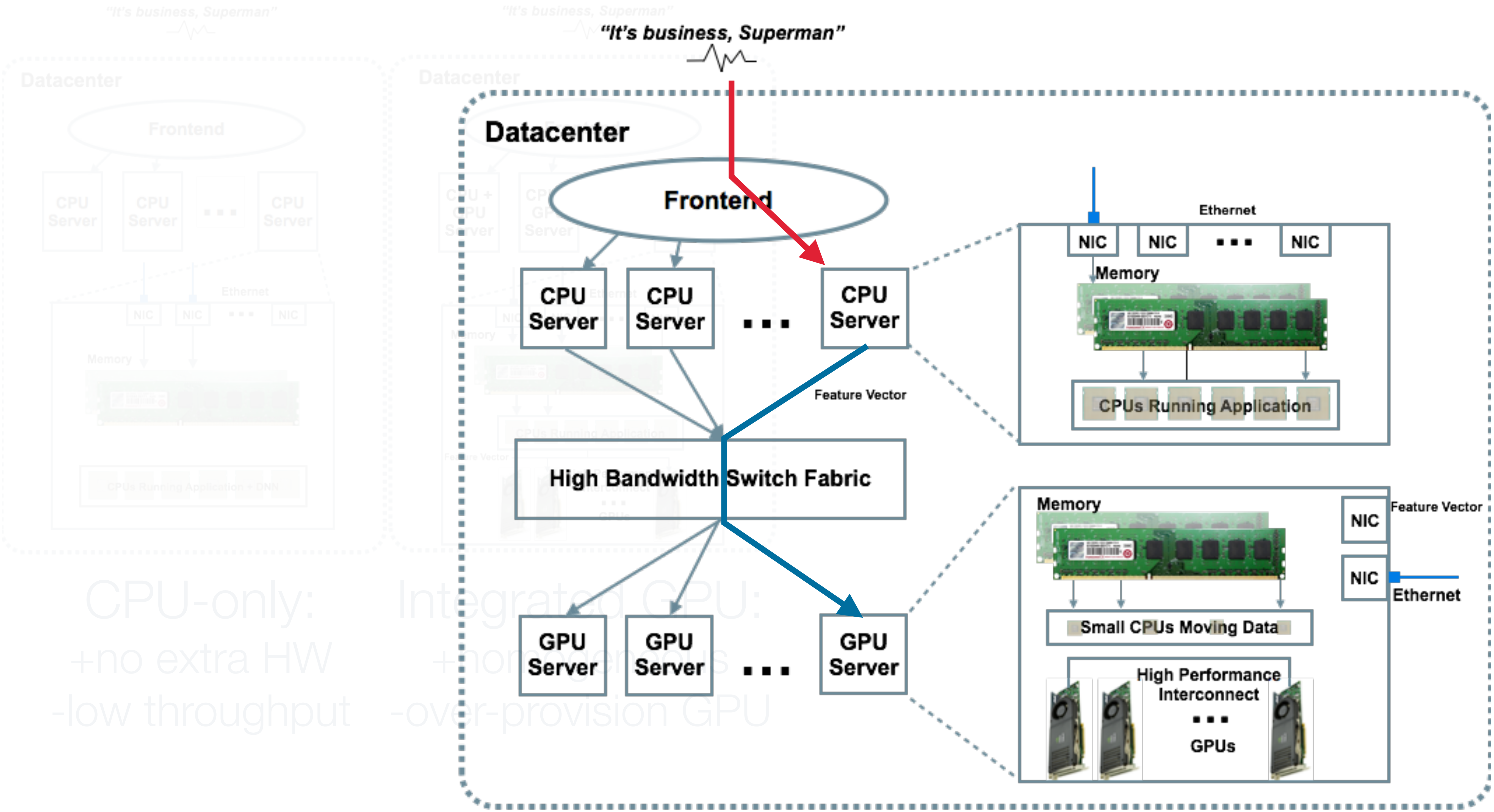
# Future WSC Designs — Server Designs



CPU-only:
+no extra HW
-low throughput

Integrated GPU:
+homogeneous
-over-provision GPU

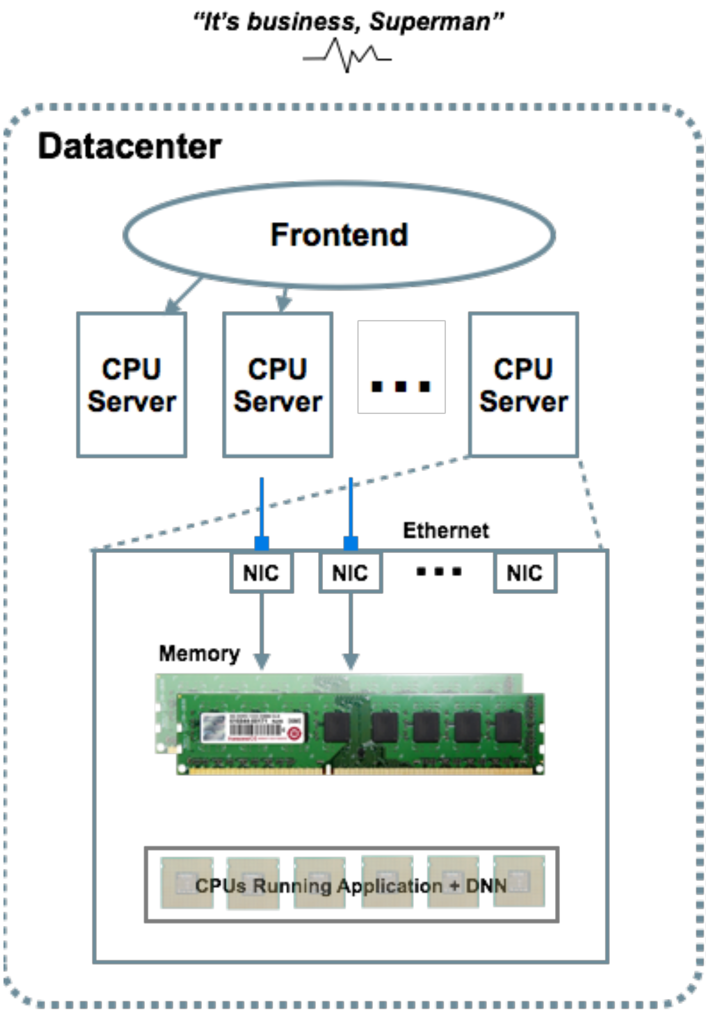
# Future WSC Designs — Server Designs



CPU-only:
+no extra HW
-low throughput

Integrated GPU:
+homogeneous
-over-provision GPU

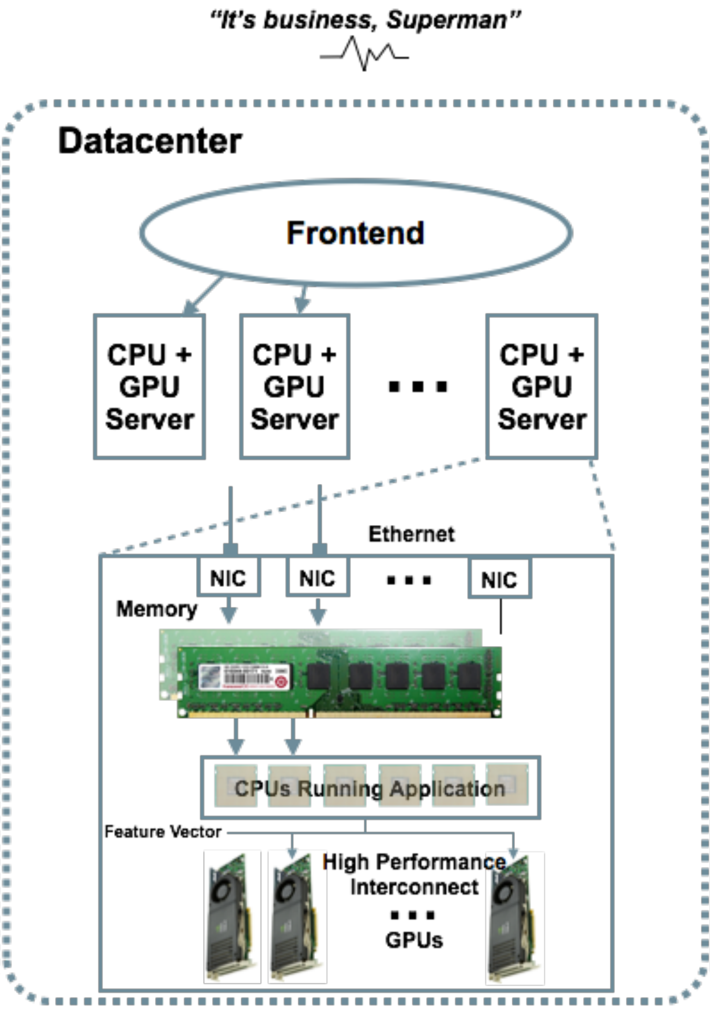
# Future WSC Designs — Server Designs



CPU-only:

+no extra HW

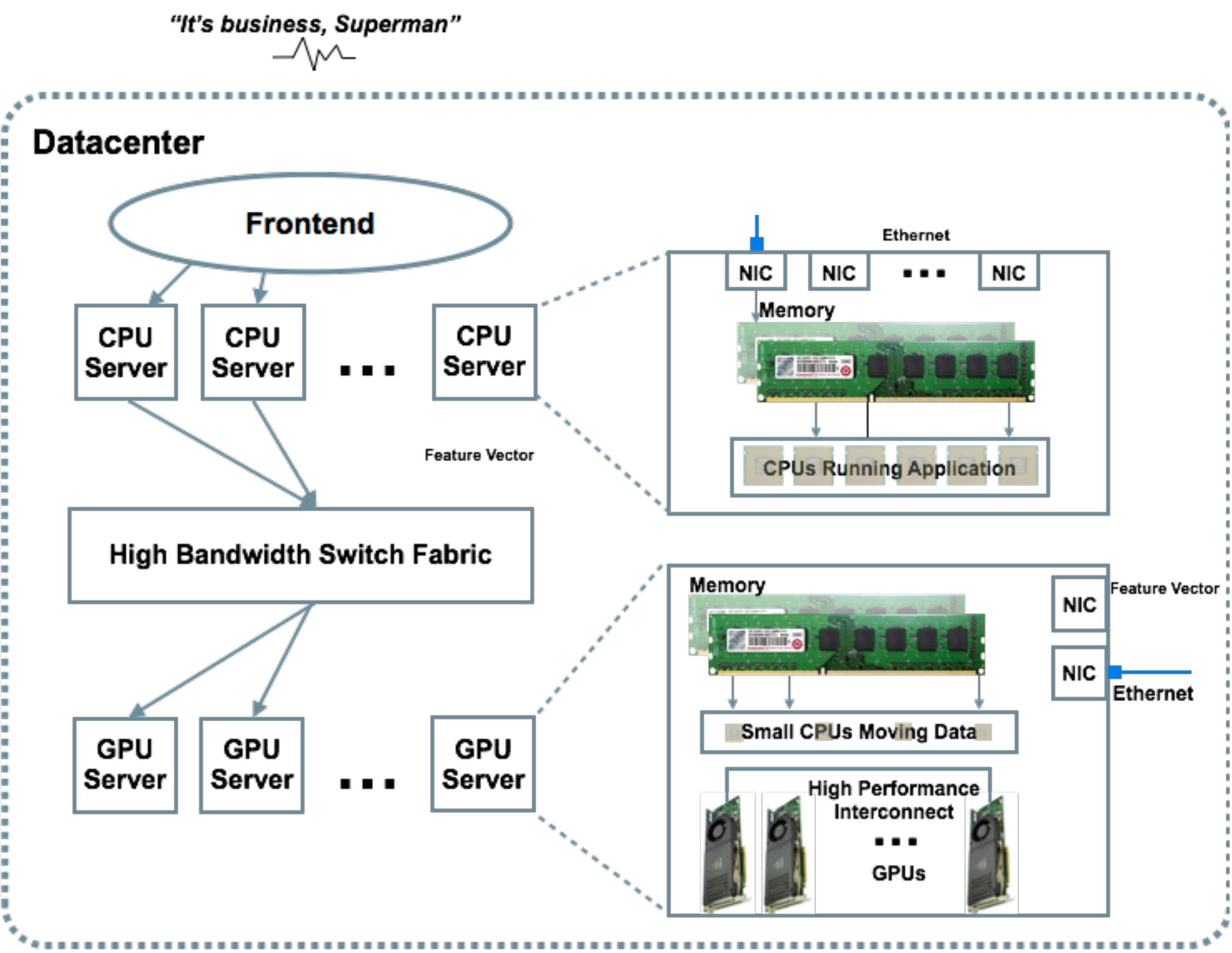
-low throughput



Integrated GPU:

+homogeneous

-over-provision GPU



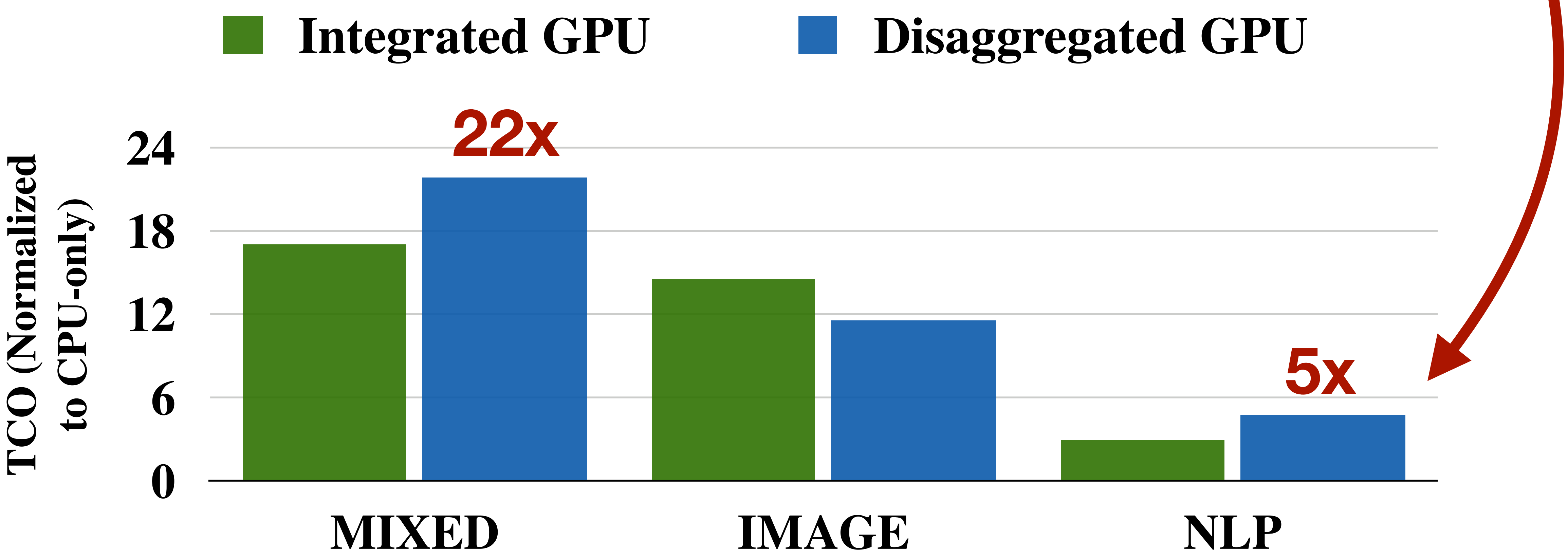
Disaggregated GPU:

+decouple CPU/GPU

-data transfer

# Future WSC Designs — Total Cost of Ownership (TCO)

- Expand Barroso [1] model with GPU and networking costs
- Addressing the bandwidth bottleneck further improves NLP TCO (more details in paper)



[1] Barroso, Luiz André, et. al. "The datacenter as a computer: An introduction to the design of warehouse-scale machines."

# WSC Scale Key Insights

- TCO improvement dependent on the DNN workload composition
  - Bandwidth constrained workloads underutilize GPUs in Integrated design
  - Sufficient bandwidth is critical to fully utilize resources

# Conclusion

- Unified, **state-of-the-art**, highly **optimized** DNN as a Service
- DNNs have **different** compute and communication characteristics
    - **Do not benefit equally** from optimizations
    - Characteristics **impact** system designs

# Thank you

DjiNN
& Tonic

**djinn.clarity-lab.org**