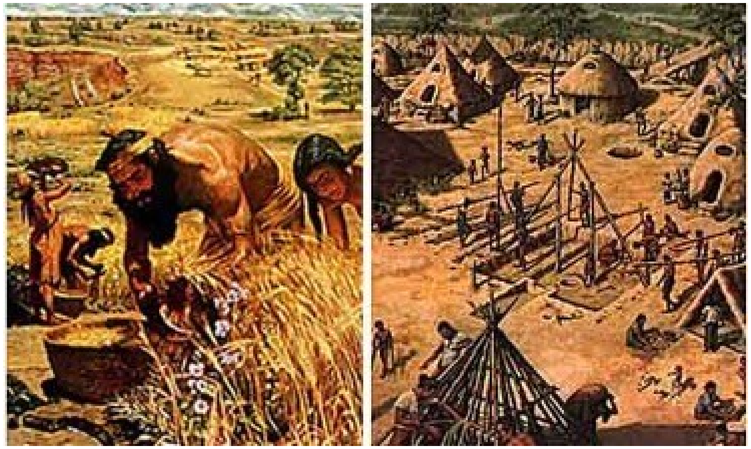
Hewlett Packard
Enterprise

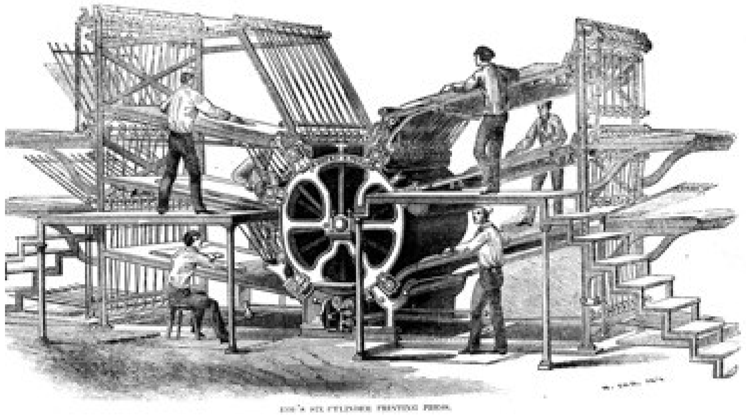
Deep Learning at HPE

Are we on the brink of a ….

Change 1:
Moving from gather and hunting to settling down to farms and ports



Change 2:
Developing the printing press and industrial revolution



Latest Change:
The greatest change of our lives. Artificial Intelligence

Where would the road take us?

Advances in artificial intelligence will transform modern life by reshaping transportation, health, science, finance, and the military.

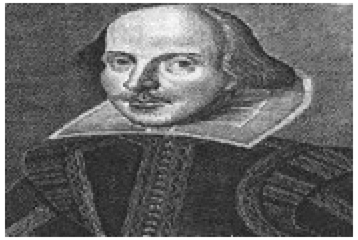
“High-level machine intelligence” (HLMI) is achieved when unaided machines can ac- complish every task better and more cheaply than human workers.

Driving a truck - 2027

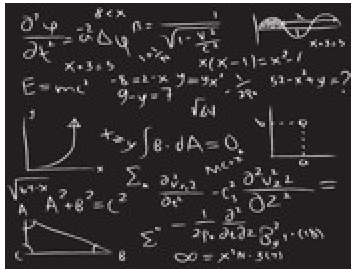
Retail - 2031

Surgeon - 2043

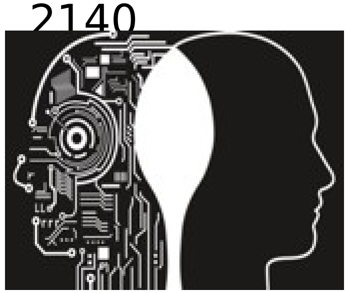
Writing a bestseller – 2049



Math Research - 2060



Full Automation of labor – 2140



Grace et al , When Will AI Exceed Human Performance? Evidence from AI Experts

The evolution of artificial intelligence

Small data sets

Early artificial intelligence

- **ENIAC** Heralded the Giant Brain“; used for post WW II ballistics

Statistical and mathematical models
applied to solve problems

1940 – 1980

Massive structured data sets

Machine learning

- **Deep Blue** Beating World Chess Champion Kasparov
- **DARPA Challenge** Autonomous vehicle drove 132 miles

Advanced Analytics and Heuristic

1990 – 2000s

Massive unstructured big data

Deep Learning

- Unsupervised training
- Generic code
- Pattern recognition

Systems can

- Observe
- Test
- Refine

Successes

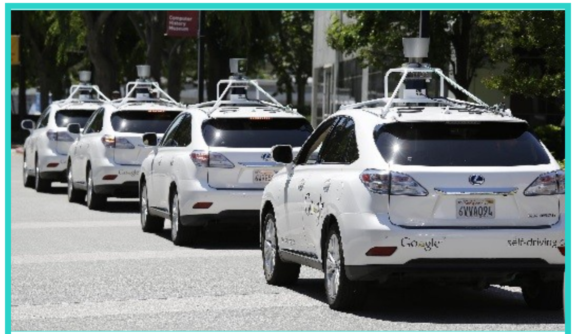
- **AlphaGO** First Computer GO program to beat a human
- **Deep Face** Facial verification
- **Libratus** AI Poker App
- **Digital virtual assistants** Siri
- **Google** Self-driving cars

Predictive models defined by
machines based neural networks

Today

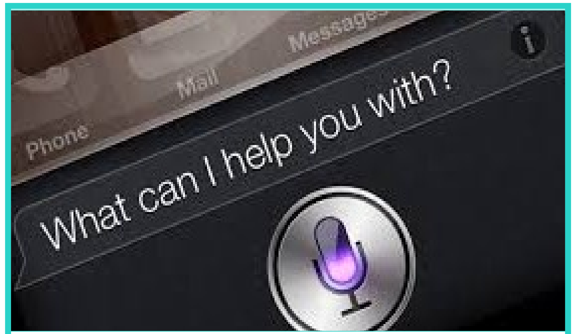
Why deep learning?

Applications



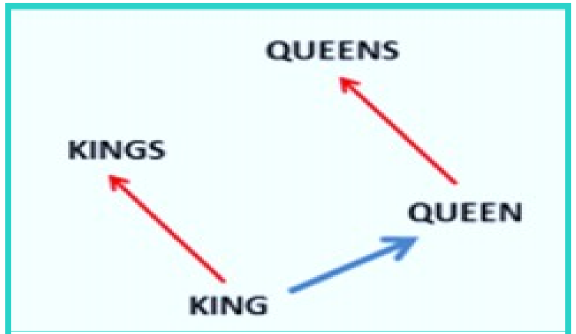
Vision

- Search & information extraction
- Security/Video surveillance
- Self-driving cars
- Medical imaging
- Robotics



Speech

- Interactive voice response (IVR) systems
- Voice interfaces (Mobile, Cars, Gaming, Home)
- Security (speaker identification)
- Health care
- People with disabilities



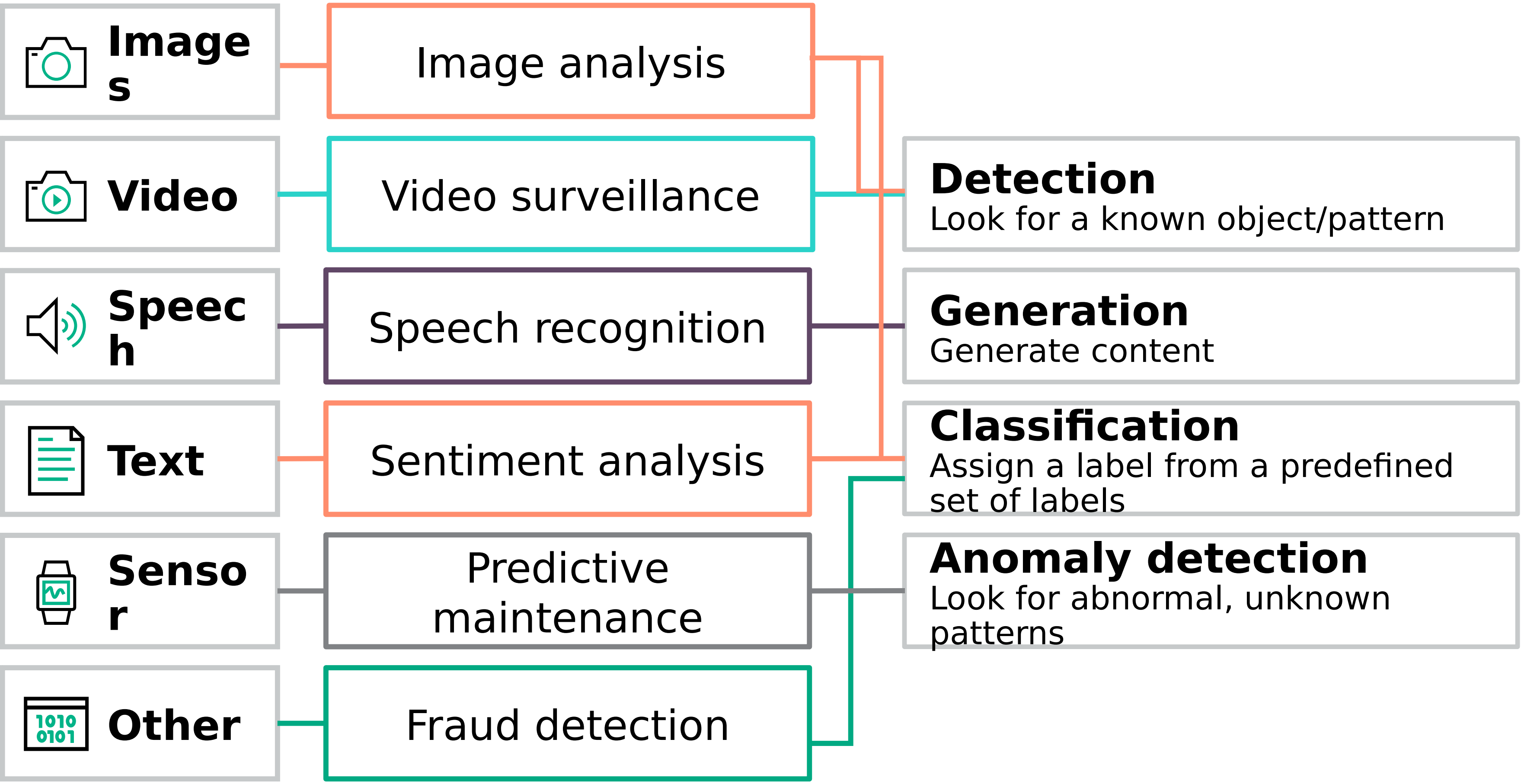
Text

- Search and ranking
- Sentiment analysis
- Machine translation
- Question answering

Other

- Recommendation engines
- Advertising
- Fraud detection
- AI challenges
- Drug discovery
- Sensor data analysis
- Diagnostic support

Applications break down



Images

Video

Speech

Text

Sensor

Other

Image analysis

Video surveillance

Speech recognition

Sentiment analysis

Predictive maintenance

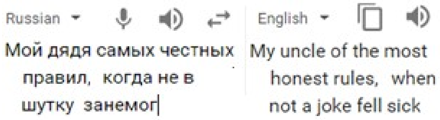
Fraud detection

Detection
Look for a known object/pattern

Generation
Generate content

Classification
Assign a label from a predefined set of labels

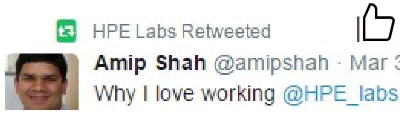
Anomaly detection
Look for abnormal, unknown patterns



Russian ▾ English ▾

Мой дядя самых честных правил, когда не в шутку занемог

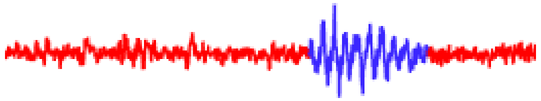
My uncle of the most honest rules, when not a joke fell sick



HPE Labs Retweeted

Amip Shah @amipshah · Mar 3

Why I love working @HPE_labs



How an individual customer’s AI evolves

Explore
How can AI help me?

Do things better
- Product development
- Customer experience
- Productivity
- Employee experience

Do new things
- New disruptions

Experiment
How can I get started?

Boundary constraints (regulations, etc.)

Data
Data model? Location?

How to **create** a model?
- Homegrown solution or open source?
- Simple ML or scalable DL?

Design
How to design and deploy the PoC?
- On-prem, cloud?
- How to think about inference

Performance
What is the best config to run?
How to tune the model to improve accuracy?

Scale up and Optimize
How can I scale and optimize?

Provisioning for inference

Infrastructure scale up
- Training
- Inference
- On-prem / cloud / hybrid

Data management
- Between edge and core
- Security
- Updates
- Regulations
- Tracing

Hewlett Packard
Enterprise

What about AI consumers ?

Do it yourself

Current wave of AI / Machine Learning is core to their business. All in-house

Google, Baidu, Facebook, Microsoft, Apple, etc.

How do I do it ?

Could benefit from better data science, machine learning, but it is not historically their core-competency

Banks, advertisers, healthcare, manufacturing, food, automotive, etc.

Not ready for an ASIC. Don’t know what they need exactly. Many still developing on CPUs. Can’t use solutions that can’t be verified or understood

I know better

Super-Experts – current wave is woefully inadequate

Government – DoD, DoE, NSA, NASA, etc.

Begging for higher performance ASICs.
Know exactly what they want to do. Strong technology pull.

Key IT challenges are constraining deep learning adoption

Limited knowledge, resources and capabilities

How to get started?

Introducing the Deep Learning Cookbook

"I need simple, infrastructure and software capabilities to rapidly and efficiently support deep learning app development."

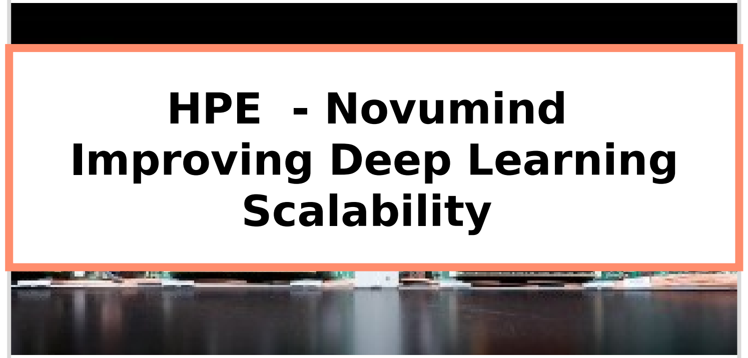
Immature, sub-optimal foundation

How to go to production?

"I could use more expert advice and tailored solutions for migrating and integrating apps in a production environment."

Inability to scale and integrate

How to scale and optimize?



HPE - Novumind Improving Deep Learning Scalability

"I need help integrating the latest technologies into my deep learning environment to accelerate actionable insights."

Lack of technology integration capabilities

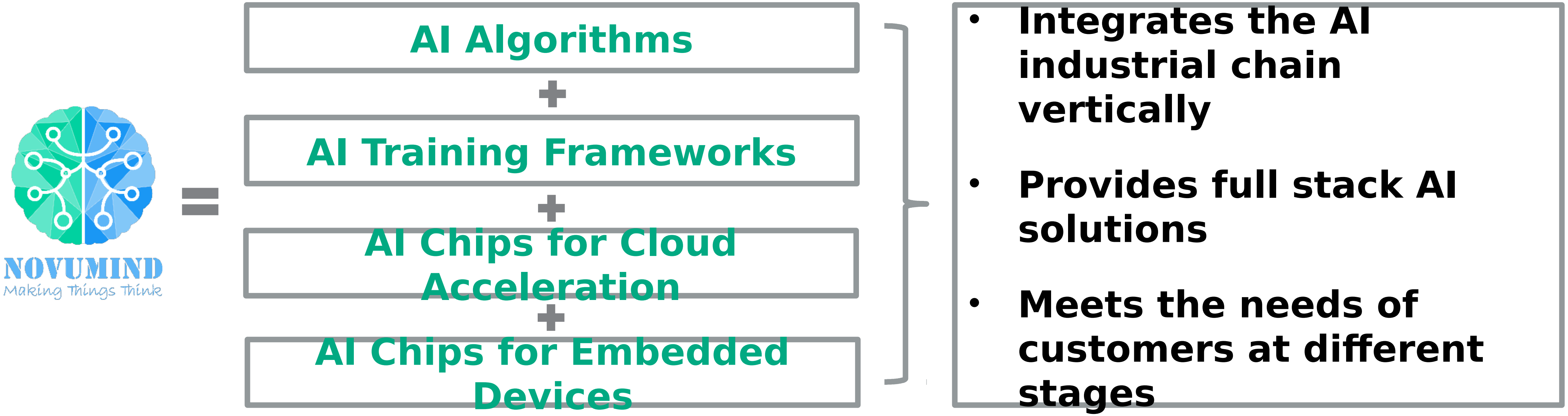
Hewlett Packard Enterprise

HPE and Novumind

NovuMind

– Providing full stack AI solutions

NovuMind is a multinational AI technology company, headquartered in the heart of Silicon Valley, with branch offices in Beijing, Hong Kong, Guangzhou and Taipei.



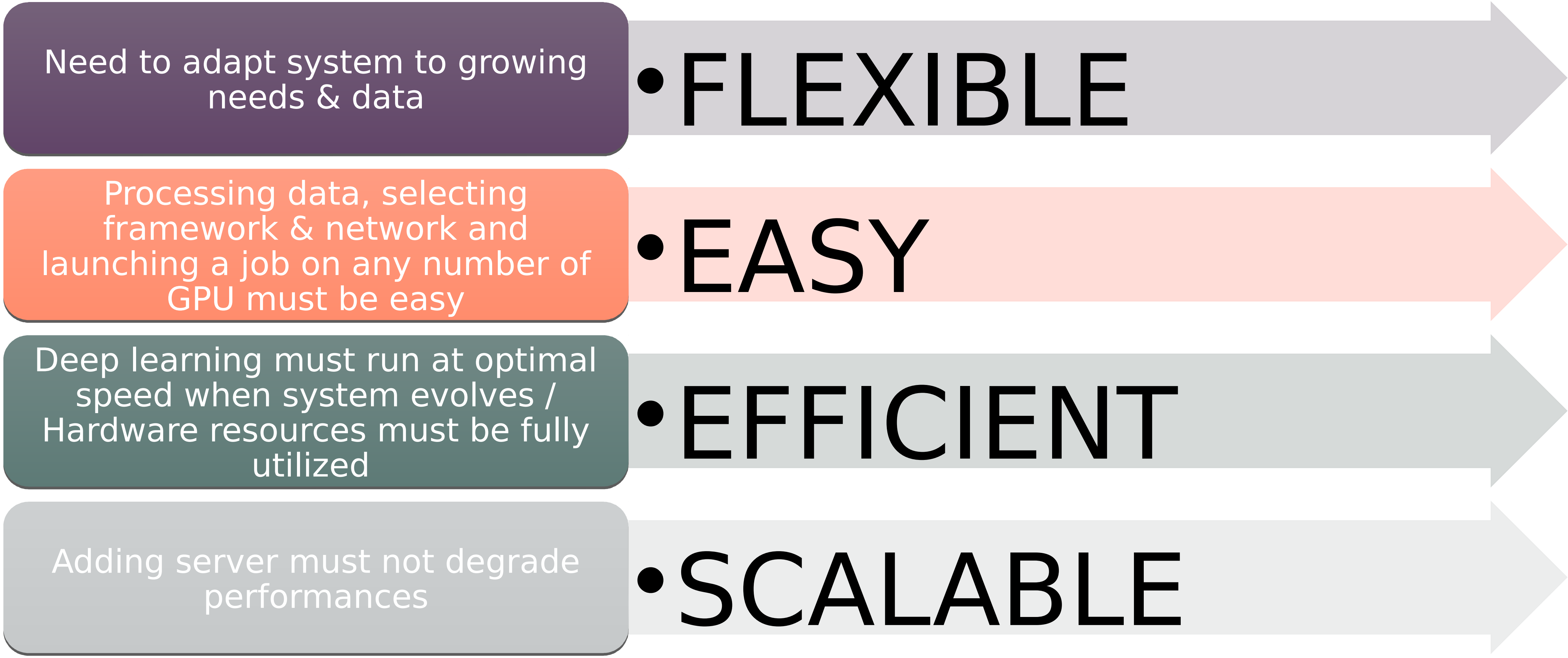
NOVUMIND
Making Things Think

=

AI Algorithms
+
AI Training Frameworks
+
AI Chips for Cloud Acceleration
+
AI Chips for Embedded Devices

- **Integrates the AI industrial chain vertically**
- **Provides full stack AI solutions**
- **Meets the needs of customers at different stages**

Which problems are we trying to solve ?



Need to adapt system to growing needs & data

- FLEXIBLE

Processing data, selecting framework & network and launching a job on any number of GPU must be easy

- EASY

Deep learning must run at optimal speed when system evolves / Hardware resources must be fully utilized

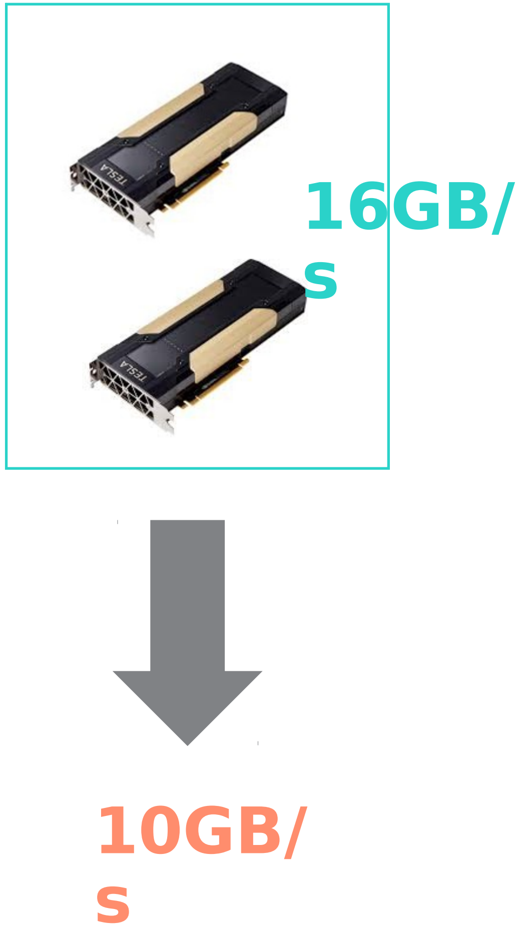
- EFFICIENT

Adding server must not degrade performances

- SCALABLE

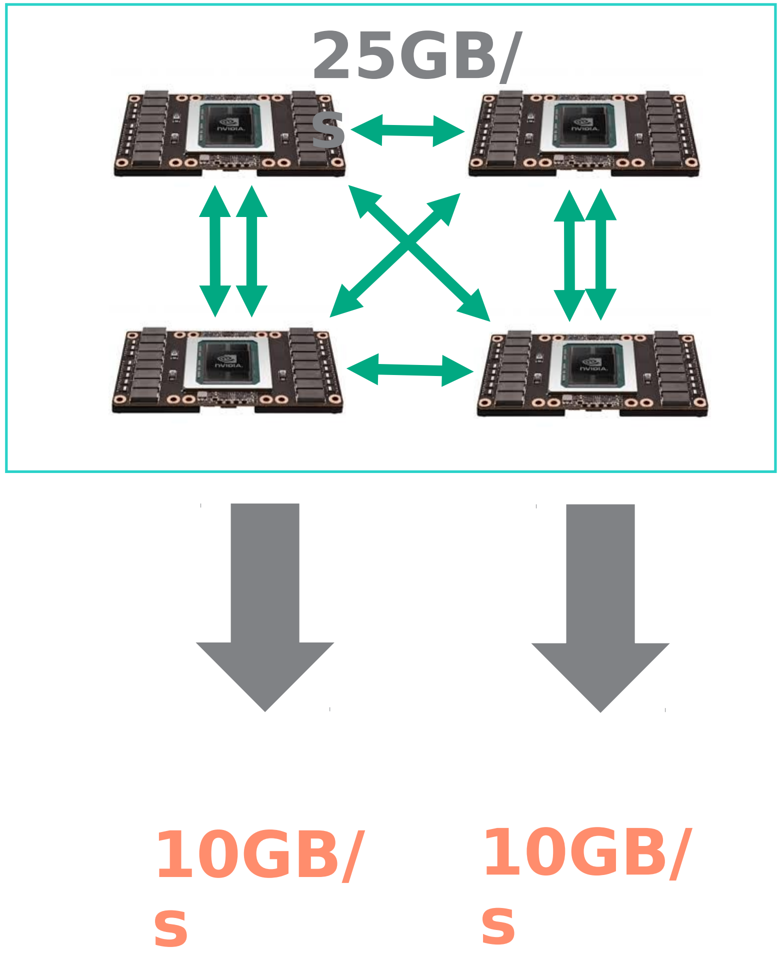
Flexibility: Topology does matter

Looking at bandwidth inter & intra node



16GB/s

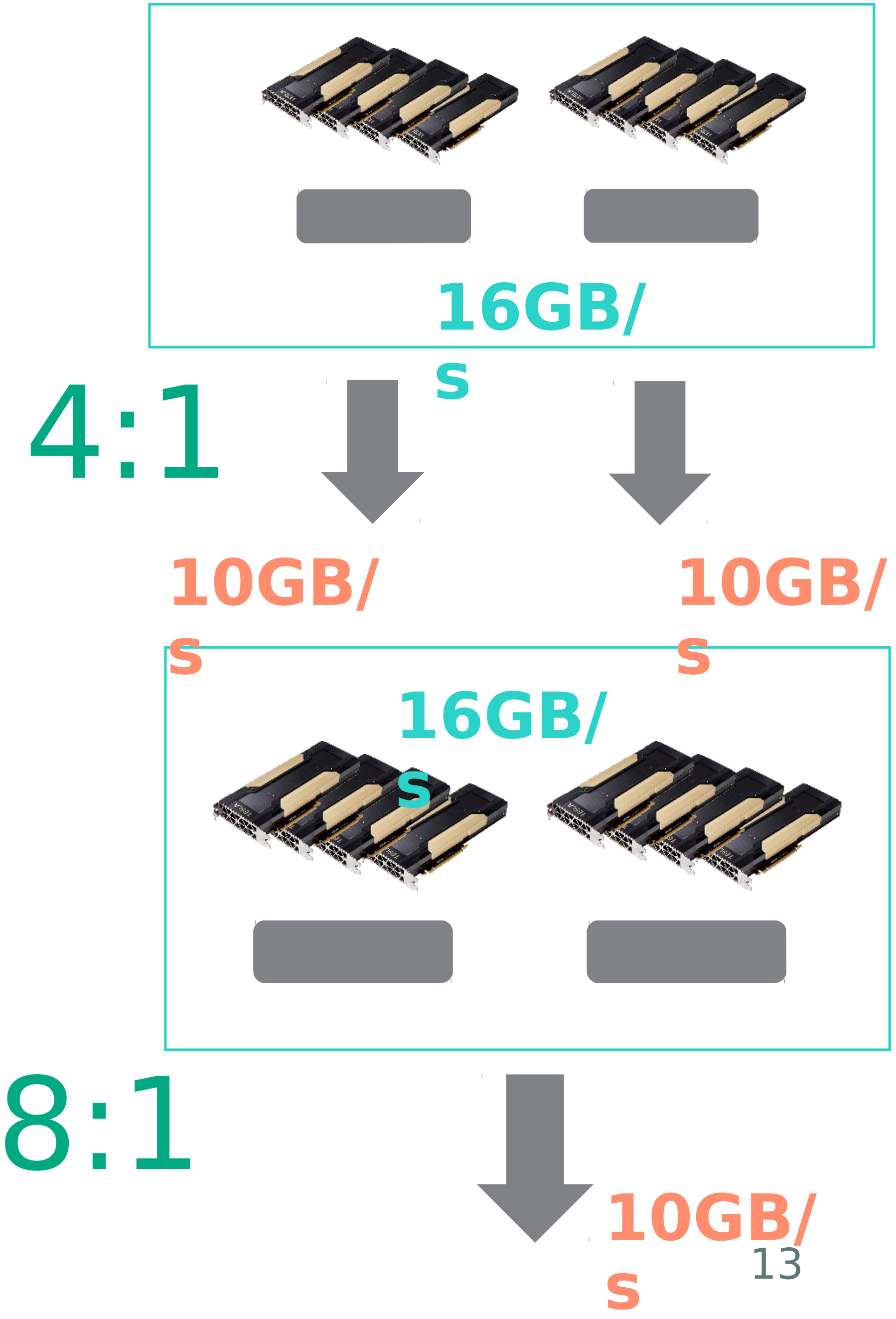
10GB/s



25GB/s

10GB/s

10GB/s



16GB/s

4:1

10GB/s

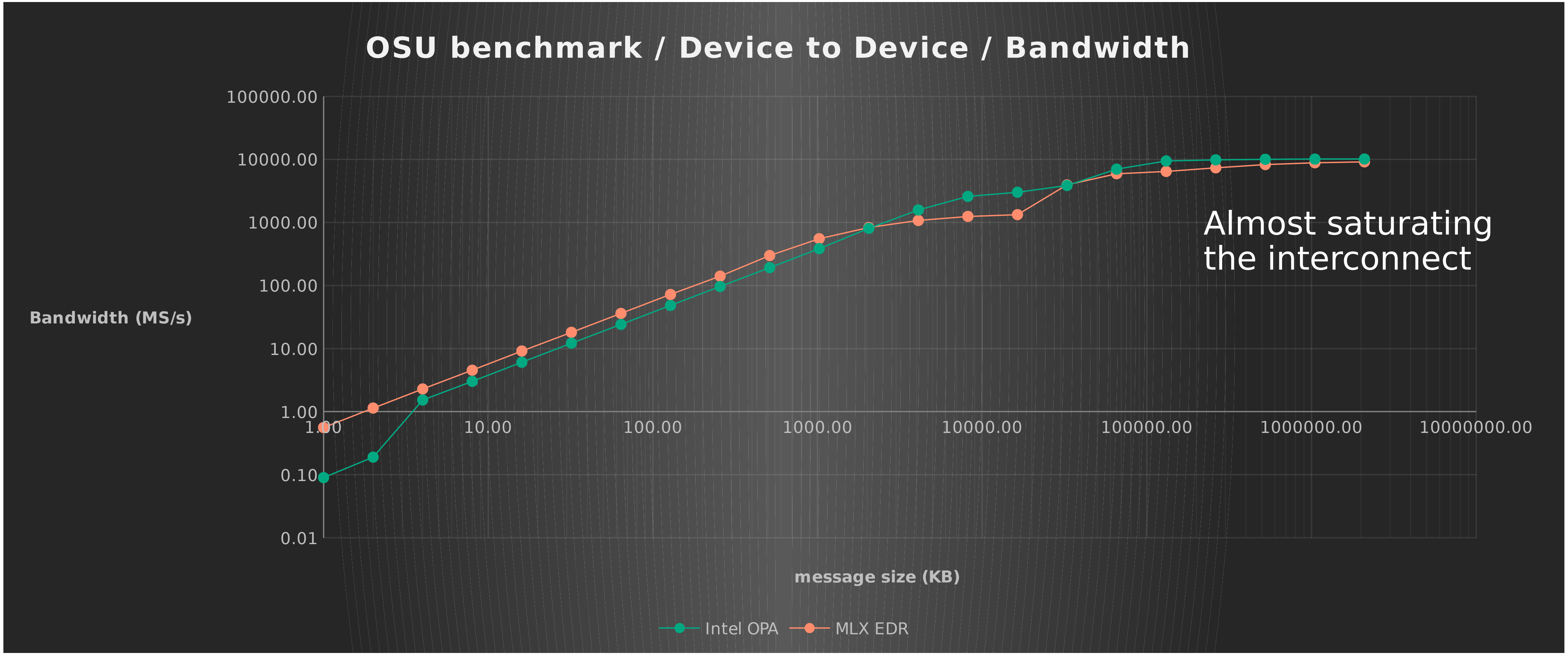
10GB/s

16GB/s

8:1

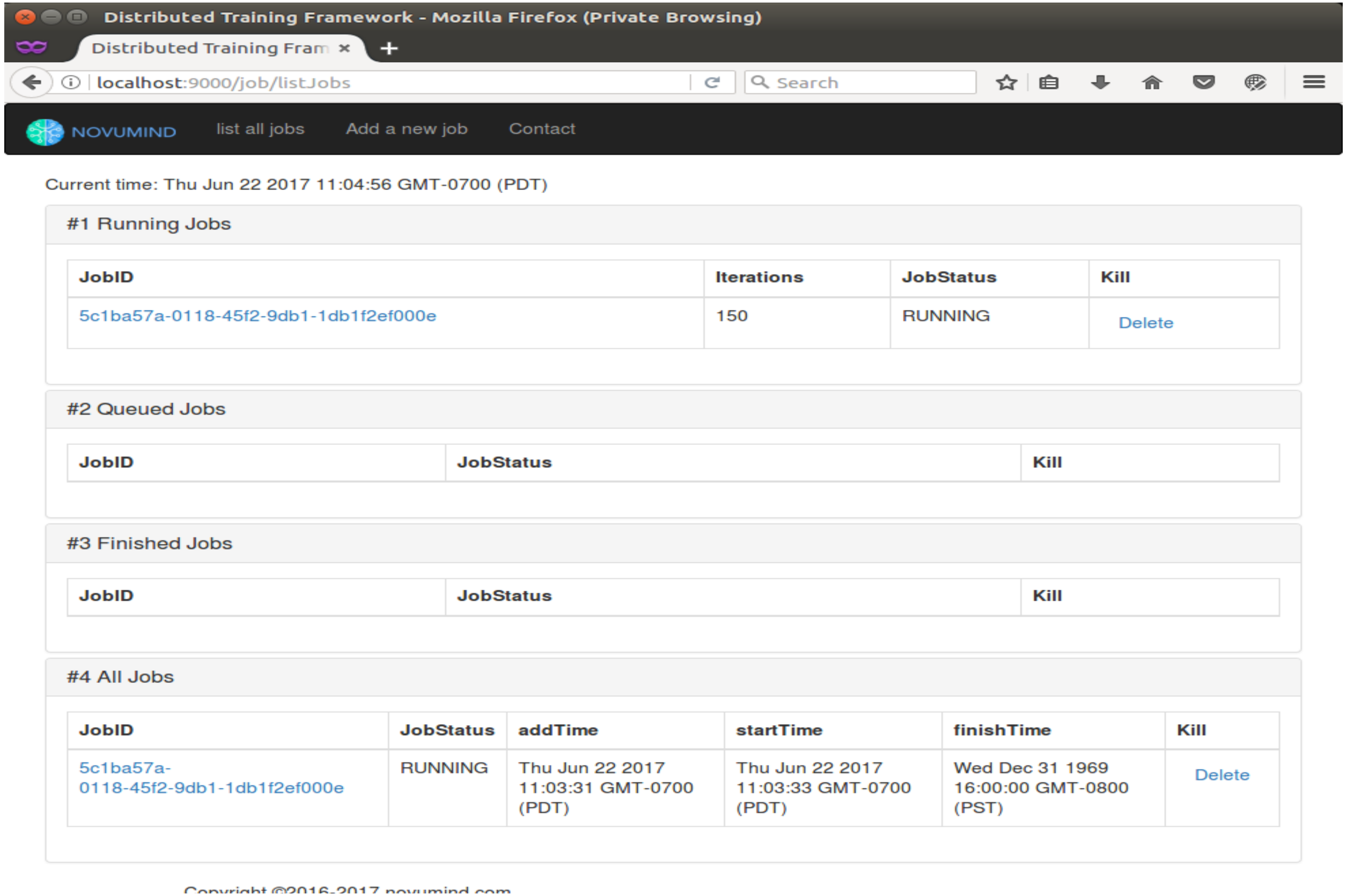
10GB/s

GPU RDMA : efficient communication to scale out



Almost saturating the interconnect

Easy: Job management Web User Interface


Distributed Training Framework - Mozilla Firefox (Private Browsing)
Distributed Training Fram ×
localhost:9000/job/listJobs
Search
NOVUMIND
list all jobs
Add a new job
Contact

Current time: Thu Jun 22 2017 11:04:56 GMT-0700 (PDT)

#1 Running Jobs

JobID	Iterations	JobStatus	Kill
5c1ba57a-0118-45f2-9db1-1db1f2ef000e	150	RUNNING	Delete

#2 Queued Jobs

JobID	JobStatus	Kill

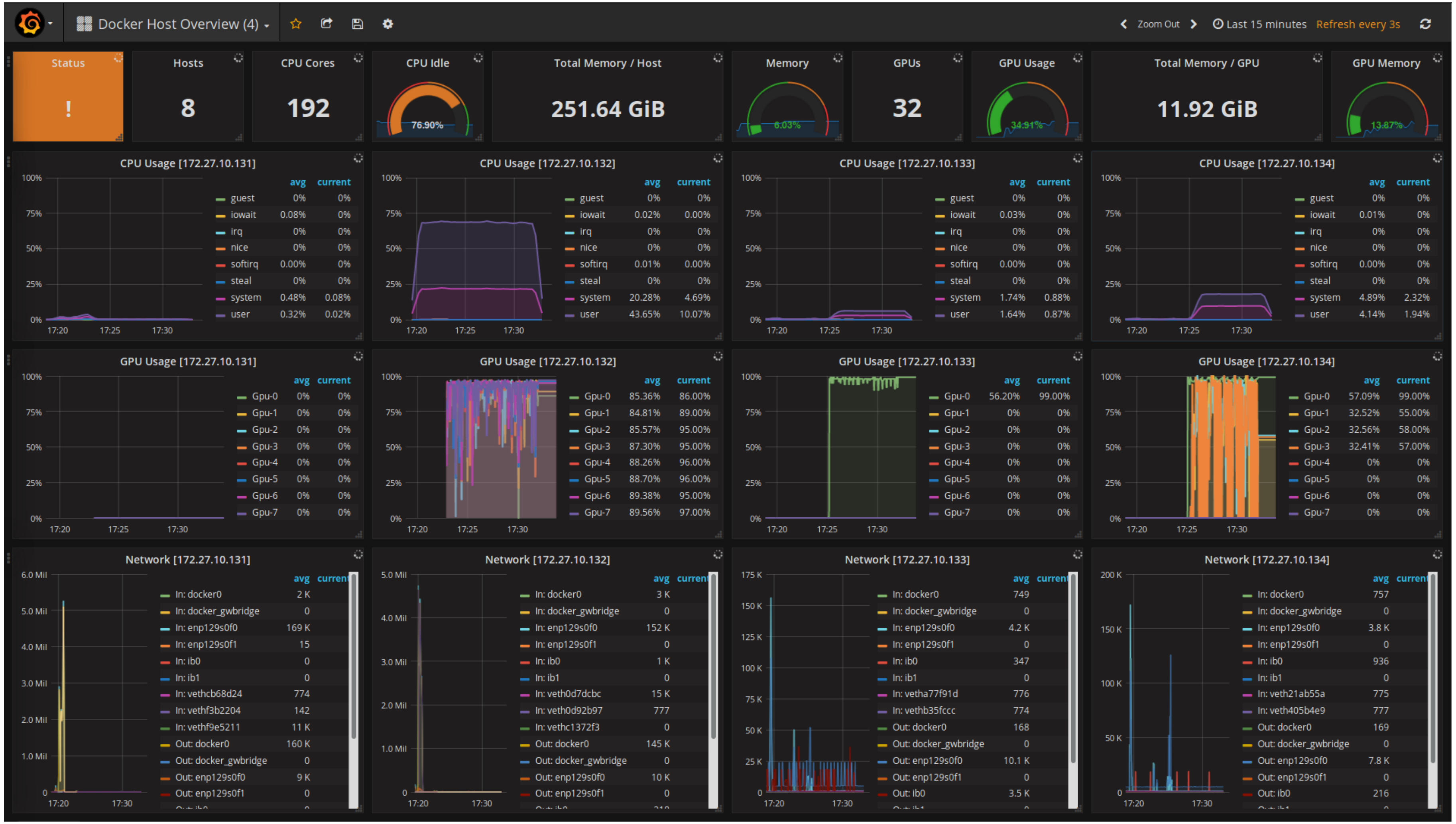
#3 Finished Jobs

JobID	JobStatus	Kill

#4 All Jobs

JobID	JobStatus	addTime	startTime	finishTime	Kill
5c1ba57a-0118-45f2-9db1-1db1f2ef000e	RUNNING	Thu Jun 22 2017 11:03:31 GMT-0700 (PDT)	Thu Jun 22 2017 11:03:33 GMT-0700 (PDT)	Wed Dec 31 1969 16:00:00 GMT-0800 (PST)	Delete

Easy: Resource Monitoring



Status	Hosts	CPU Cores	CPU Idle	Total Memory / Host	Memory	GPUs	GPU Usage	Total Memory / GPU	GPU Memory
!	8	192	76.90%	251.64 GiB	6.03%	32	34.91%	11.92 GiB	13.87%

CPU Usage [172.27.10.131]

	avg	current
guest	0%	0%
iowait	0.08%	0%
irq	0%	0%
nice	0%	0%
softirq	0.00%	0%
steal	0%	0%
system	0.48%	0.08%
user	0.32%	0.02%

CPU Usage [172.27.10.132]

	avg	current
guest	0%	0%
iowait	0.02%	0.00%
irq	0%	0%
nice	0%	0%
softirq	0.01%	0.00%
steal	0%	0%
system	20.28%	4.69%
user	43.65%	10.07%

CPU Usage [172.27.10.133]

	avg	current
guest	0%	0%
iowait	0.03%	0%
irq	0%	0%
nice	0%	0%
softirq	0.00%	0%
steal	0%	0%
system	1.74%	0.88%
user	1.64%	0.87%

CPU Usage [172.27.10.134]

	avg	current
guest	0%	0%
iowait	0.01%	0%
irq	0%	0%
nice	0%	0%
softirq	0.00%	0%
steal	0%	0%
system	4.89%	2.32%
user	4.14%	1.94%

GPU Usage [172.27.10.131]

	avg	current
Gpu-0	0%	0%
Gpu-1	0%	0%
Gpu-2	0%	0%
Gpu-3	0%	0%
Gpu-4	0%	0%
Gpu-5	0%	0%
Gpu-6	0%	0%
Gpu-7	0%	0%

GPU Usage [172.27.10.132]

	avg	current
Gpu-0	85.36%	86.00%
Gpu-1	84.81%	89.00%
Gpu-2	85.57%	95.00%
Gpu-3	87.30%	95.00%
Gpu-4	88.26%	96.00%
Gpu-5	88.70%	96.00%
Gpu-6	89.38%	95.00%
Gpu-7	89.56%	97.00%

GPU Usage [172.27.10.133]

	avg	current
Gpu-0	56.20%	99.00%
Gpu-1	0%	0%
Gpu-2	0%	0%
Gpu-3	0%	0%
Gpu-4	0%	0%
Gpu-5	0%	0%
Gpu-6	0%	0%
Gpu-7	0%	0%

GPU Usage [172.27.10.134]

	avg	current
Gpu-0	57.09%	99.00%
Gpu-1	32.52%	55.00%
Gpu-2	32.56%	58.00%
Gpu-3	32.41%	57.00%
Gpu-4	0%	0%
Gpu-5	0%	0%
Gpu-6	0%	0%
Gpu-7	0%	0%

Network [172.27.10.131]

	avg	current
In: docker0	2 K	
In: docker_gwbridge	0	
In: enp129s0f0	169 K	
In: enp129s0f1	15	
In: ib0	0	
In: ib1	0	
In: vethcb68d24	774	
In: vethf3b2204	142	
In: vethf9e5211	11 K	
Out: docker0	160 K	
Out: docker_gwbridge	0	
Out: enp129s0f0	9 K	
Out: enp129s0f1	0	

Network [172.27.10.132]

	avg	current
In: docker0	3 K	
In: docker_gwbridge	0	
In: enp129s0f0	152 K	
In: enp129s0f1	0	
In: ib0	1 K	
In: ib1	0	
In: veth0d7dcbc	15 K	
In: veth0d92b97	777	
In: vethc1372f3	0	
Out: docker0	145 K	
Out: docker_gwbridge	0	
Out: enp129s0f0	10 K	
Out: enp129s0f1	0	

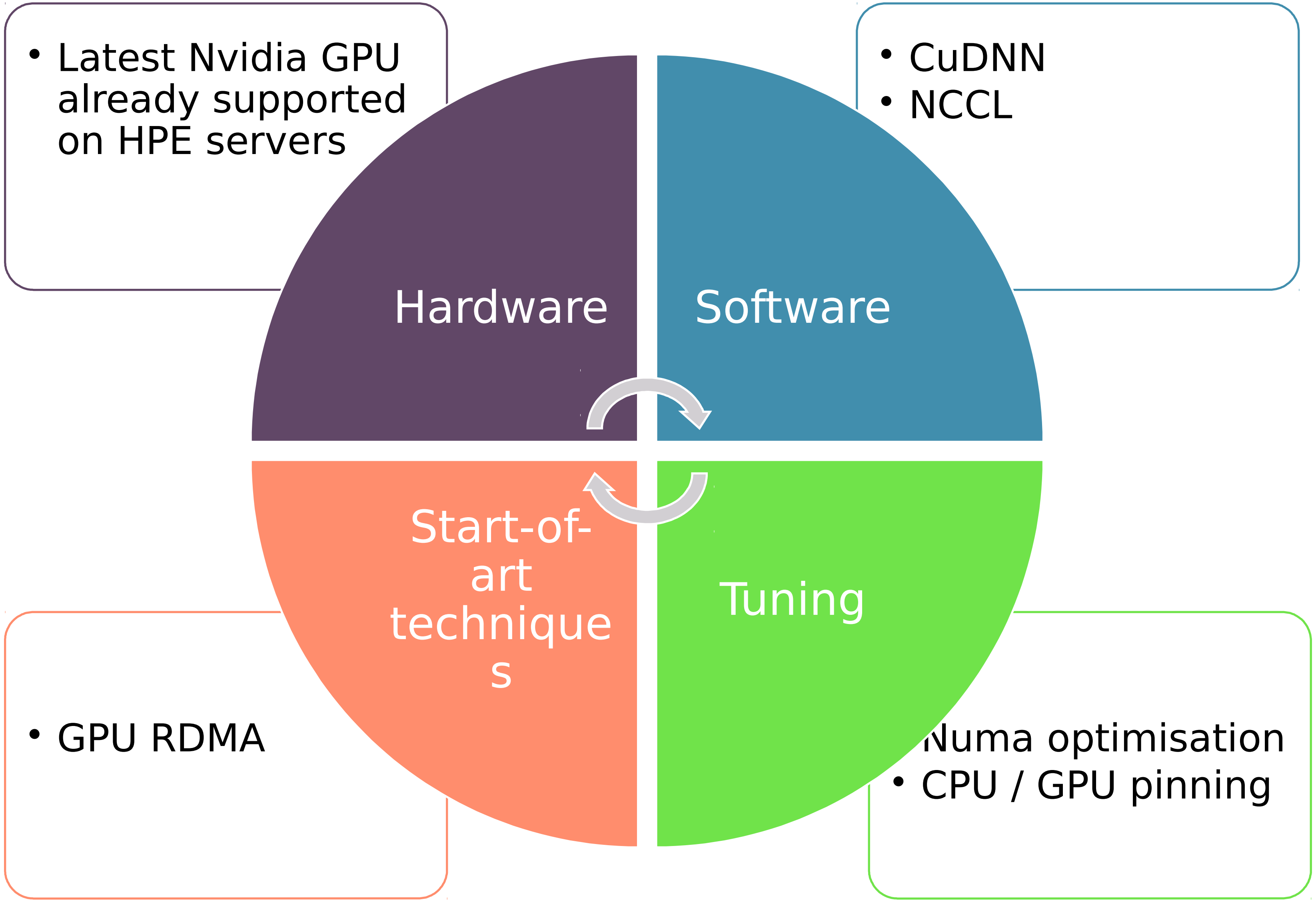
Network [172.27.10.133]

	avg	current
In: docker0	749	
In: docker_gwbridge	0	
In: enp129s0f0	4.2 K	
In: enp129s0f1	0	
In: ib0	347	
In: ib1	0	
In: vetha77f91d	776	
In: vethb35fccc	774	
Out: docker0	168	
Out: docker_gwbridge	0	
Out: enp129s0f0	10.1 K	
Out: enp129s0f1	0	
Out: ib0	3.5 K	

Network [172.27.10.134]

	avg	current
In: docker0	757	
In: docker_gwbridge	0	
In: enp129s0f0	3.8 K	
In: enp129s0f1	0	
In: ib0	936	
In: ib1	0	
In: veth21ab55a	775	
In: veth405b4e9	777	
Out: docker0	169	
Out: docker_gwbridge	0	
Out: enp129s0f0	7.8 K	
Out: enp129s0f1	0	
Out: ib0	216	

Efficient



Hardware

- Latest Nvidia GPU already supported on HPE servers

Software

- CuDNN
- NCCL

Start-of-art techniques

- GPU RDMA

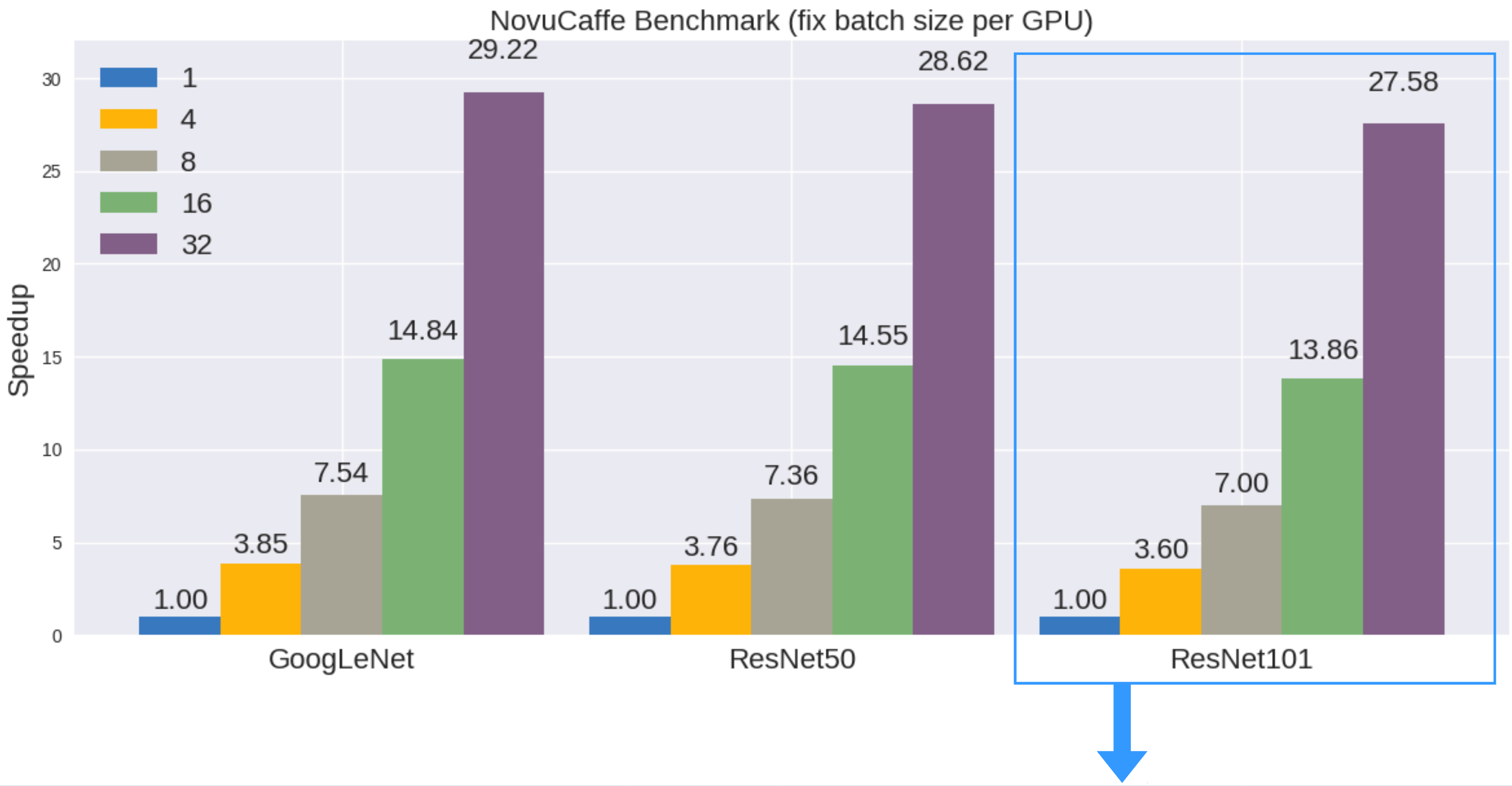
Tuning

- Numa optimisation
- CPU / GPU pinning

Efficient: Leveraging Novumind experience for optimal runs

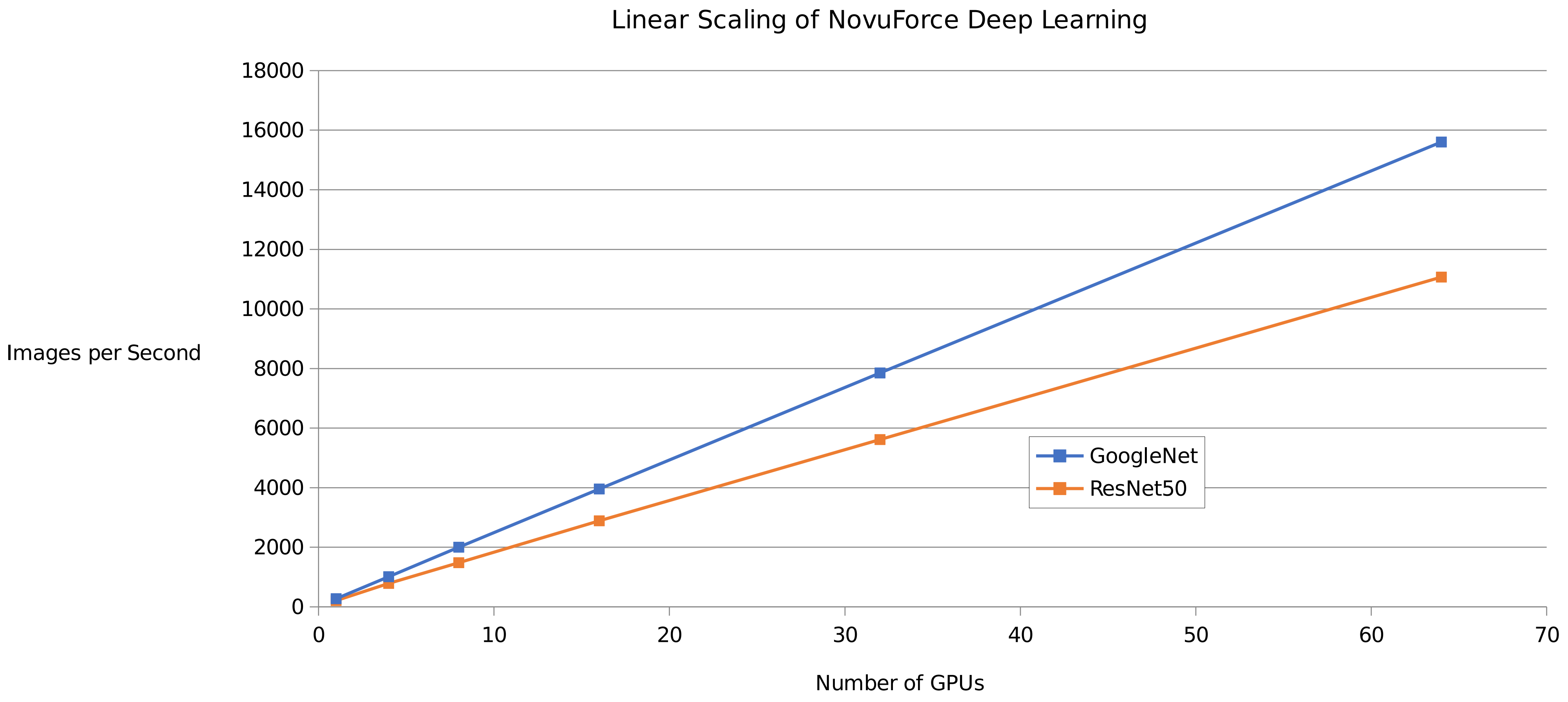
- Data Augmentation
  - scale and aspect
  - Color
  - Weight decay
- Per epoch data shuffling
- Base on Novumind's domain experience of predefined set of meta-parameter, tuning become cook book recipes.
- In certain scenario, LR scarified a little bit accuracy for much faster convergency
- Expert knowledge of past experience to tune neural networks. For example, in Image classification, No need to search for potential tuning. Novuforce will suggest optimal ones.
- Different verticals optimal parameters (security, heathcare, transportation, financial services)

Scalability: Benchmark Results



Iterations	Model	Batch Size	GPUs	Elapsed (s)	Speedups	Final Loss
2900	resnet101	1 * 128	1	12027	**-**	**4.6682**
2900	resnet101	4 * 128	4	13378	**3.5961**	**3.4406**
2900	resnet101	8 * 128	8	13748	**6.9985**	**3.0311**
2900	resnet101	16 * 128	16	13889	**13.855**	**2.3069**
2900	resnet101	32 * 128	32	13954	**27.5809**	**1.7068**

Scalability: Benchmark Results



Hewlett Packard Enterprise

Deep Learning Cookbook

Where to start ?

Recommend DL stack by vertical application

Verticals	Voice interfaces	Social media	Manufacturing	Oil & gas	Connected cars
Data type	Speech	Images	Video	Sensor data	
Data	Small	Moderate	Large		
Typical layers	Convolutional	Fully-connected	Recurrent	…	
Frameworks	TensorFlow	Caffe 2	CNTK	Torch	…
Infrastructure	x86	GPUs	FPGAs	TPU ?	…

← **Neural Network sits here**

Neural Network : Popular Networks

Network	Model size (# params)	Model size (MB)	GFLOPs (forward pass)
AlexNet	60,965,224	233 MB	0.7
GoogleNet	6,998,552	27 MB	1.6
VGG-16	138,357,544	528 MB	15.5
VGG-19	143,667,240	548 MB	19.6
ResNet50	25,610,269	98 MB	3.9
ResNet101	44,654,608	170 MB	7.6
ResNet152	60,344,387	230 MB	11.3

Hardware : Today’s scale and needs

Model size, data size, compute requirements

Application	Model	Training data	FLOPs per epoch
Vision	$1.7 * 10^9$ ~6.8 GB	$14*10^6$ images ~2.5 TB (256x256) ~10 TB (512x512)	$6*1.7*10^9*14*10^6$ $\sim 1.4*10^{17}$

1 epoch per hour:
~39 TFLOPS

Today’s hardware:

NVIDIA Tesla V100: 15 TFLOPS SP (30 TFLOPS FP16 , 120 TFLOPS Tensor ops), 12 GB memory

NVIDIA Tesla P100: 10.6 TFLOPS SP, 16 GB memory

NVIDIA Tesla K40: 4.29 TFLOPS SP, 12 GB memory

NVIDIA Tesla K80: 5.6 TFLOPS SP (8.74 TFLOPS SP with GPU boost), 24 GB memory

INTEL Xeon Phi: 2.4 TFLOPS SP

Superdome X: ~21 TFLOPS SP, 24 TB memory

So what to recommend?

Software

Caffe
theano
torch
KALDI
Caffe2
Microsoft CNTK
mxnet
TensorFlow
Chainer

Hardware

NVIDIA
intel
Movidius
nervana SYSTEMS
ARM
Google
XILINX
AMD
QUALCOMM

Building performance models

Alex Net

GoogleNet

VGG-16, VGG -19

ResNet 50, 101,152

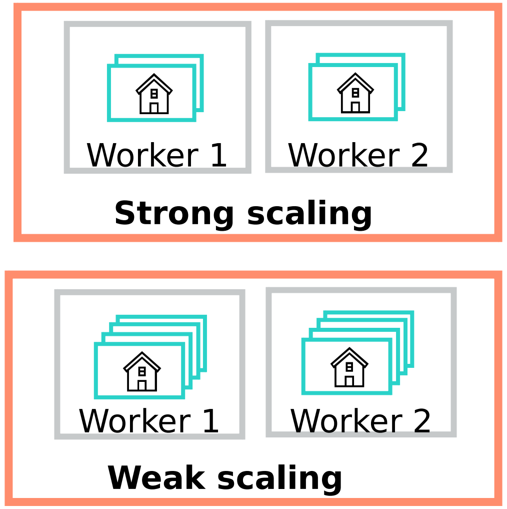
Eng Acoustic Model

TensorFlow

Caffe 2

Tensor RT

BVLC Caffe

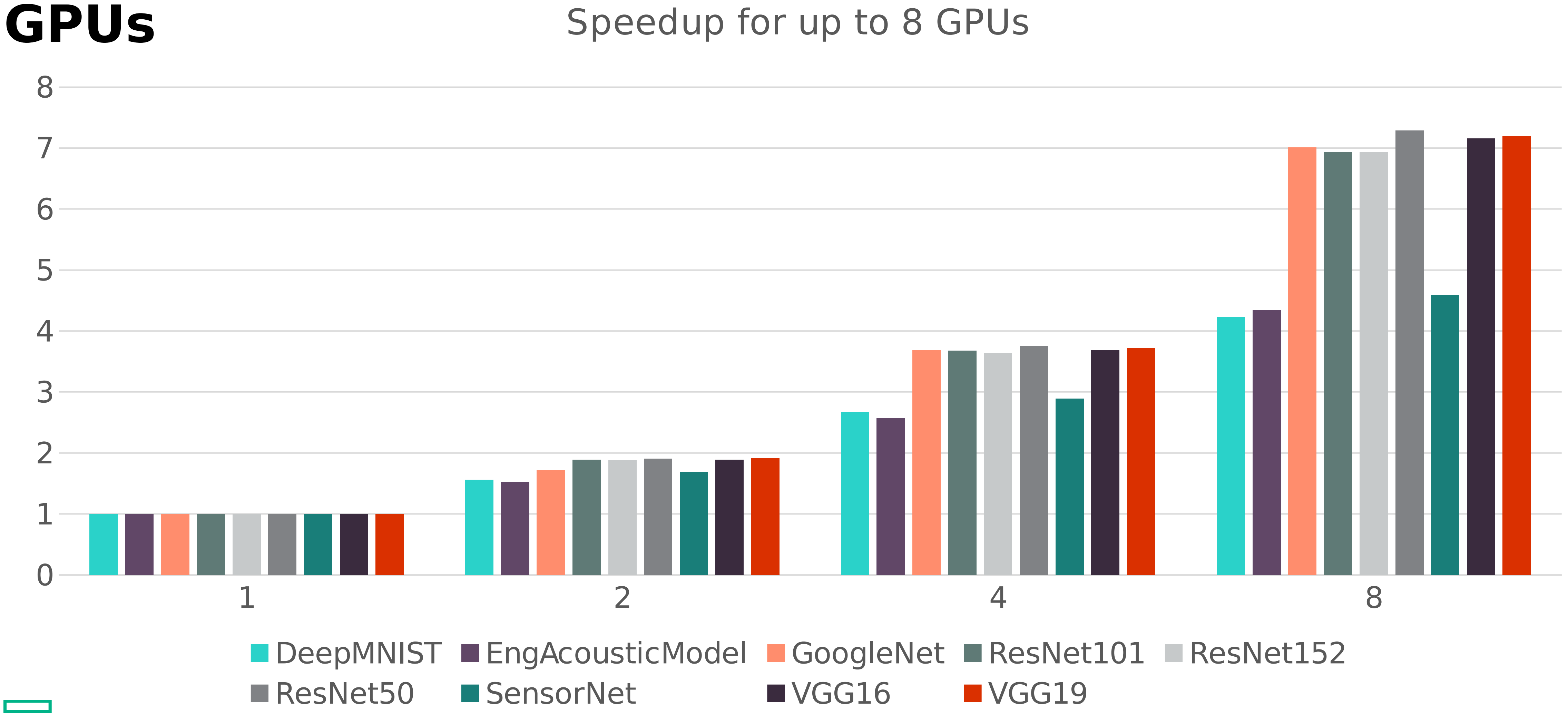


Hardware

Scalable, automated real-time intelligence

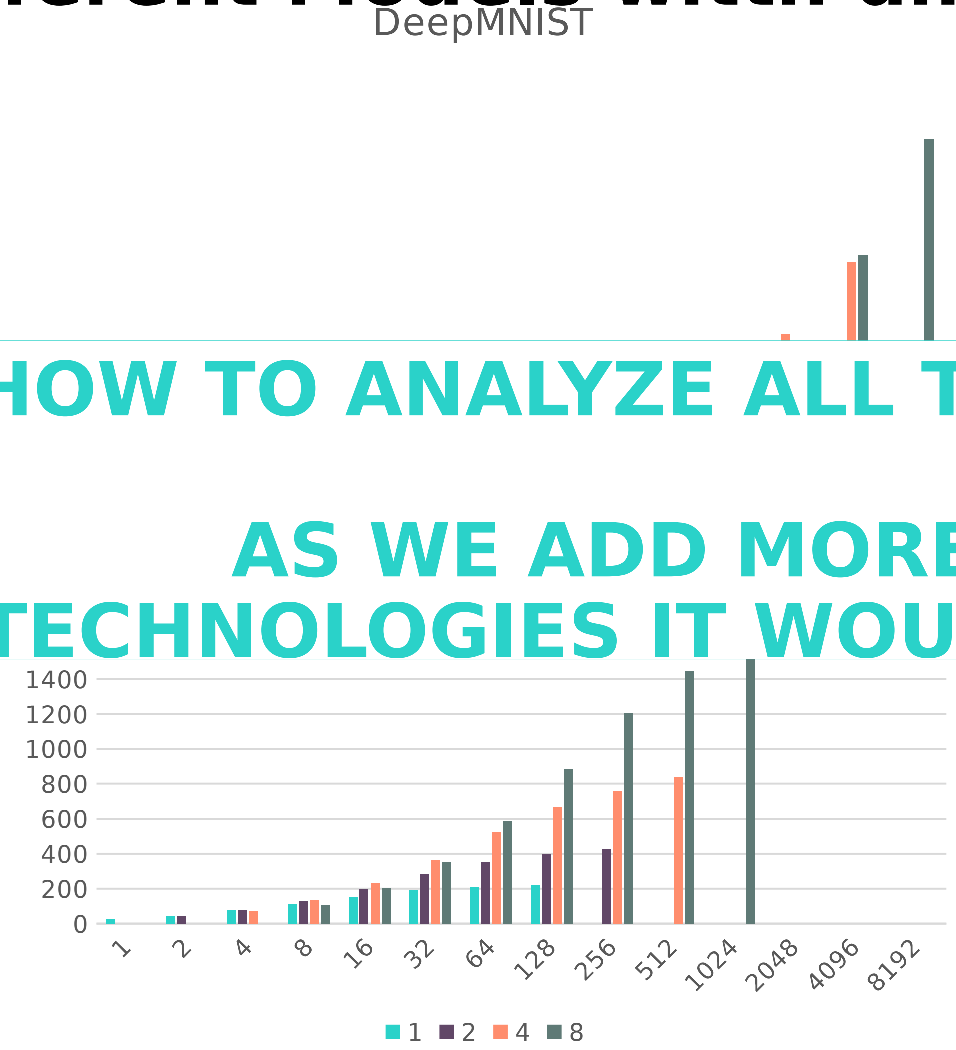
Populated with 8 GPUs

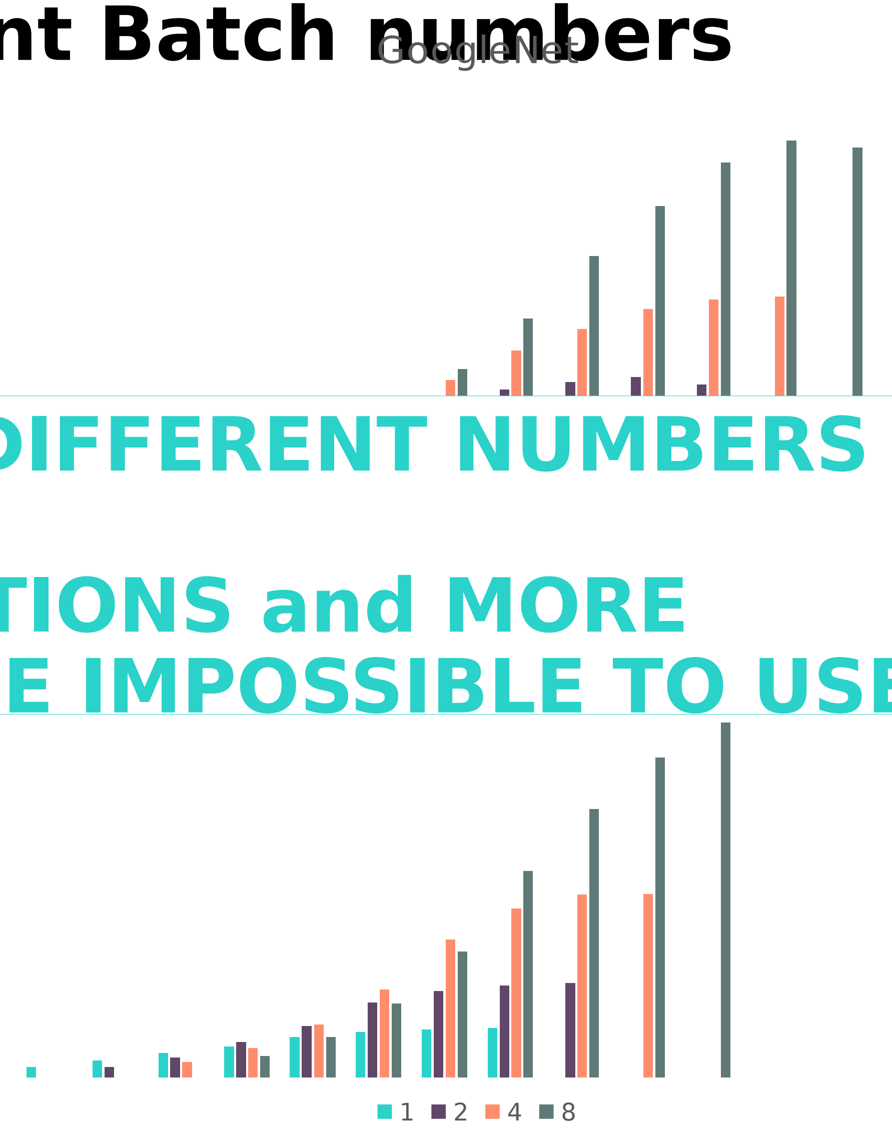
TensorFlow – Weak Scaling – Training – Different models perfromance in Tensor Flow . Scaling up to 8 GPUs



Hewlett Packard Enterprise

TensorFlow - Inference (Inferences per Second) - Different Models witth different Batch numbers





HOW TO ANALYZE ALL THE DIFFERENT NUMBERS .

AS WE ADD MORE OPTIONS and MORE TECHNOLOGIES IT WOULD BE IMPOSSIBLE TO USE

Hewlett Packard
Enterprise

Select ideal technology configurations

with HPE Deep Learning Cookbook

“Book of recipes” for deep learning workloads

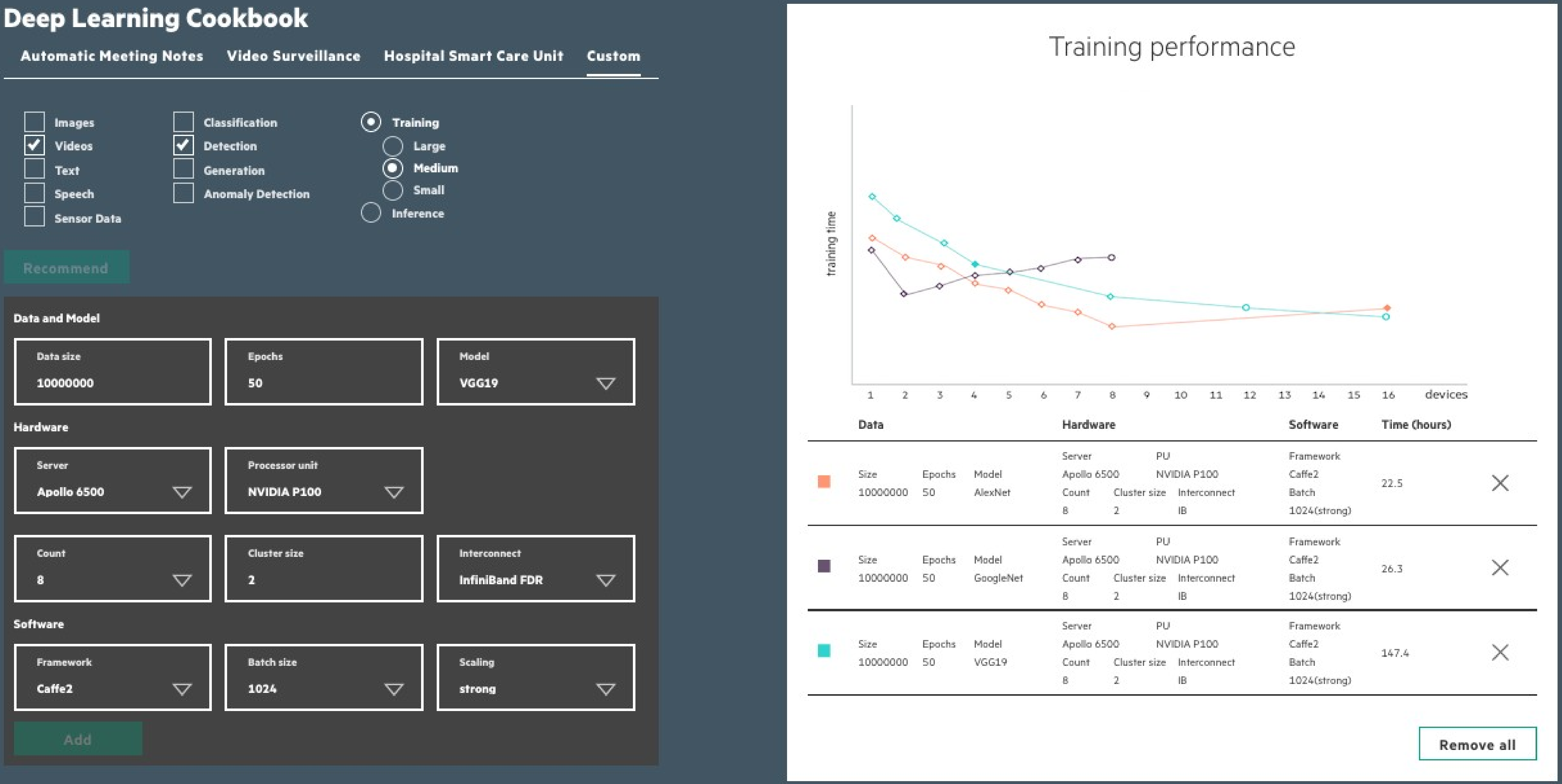
- **Comprehensive tool set** based on extensive benchmarking
- **Includes** various models with 8 DL frameworks and 8 HPE hardware systems
- **Estimates workload performance** and recommends an optimal HW/SW stack for that workload

Expert advice to get you started

- **Informed decision making** - optimal hardware and software configurations
- **Eliminates the “guesswork”** - validated methodology and data
- **Improves efficiency** - detects bottlenecks in deep learning workloads
- **Determine scalability**

Availability of complete toolset

- **Deep Learning Benchmarking Suite:** available on GitHub Dec 2018
- **Deep Learning Performance Analysis Tool:** planned to be released in the beginning of 2018**.**
- **Reference configurations:** available soon on HPE.com website



Deep Learning Cookbook

Automatic Meeting Notes | Video Surveillance | Hospital Smart Care Unit | **Custom**

□ Images
☑ Videos
□ Text
□ Speech
□ Sensor Data

□ Classification
☑ Detection
□ Generation
□ Anomaly Detection

● Training
○ Large
● Medium
○ Small
○ Inference

Recommend

Data and Model

Data size 10000000	Epochs 50	Model VGG19 ▽

Hardware

Server Apollo 6500 ▽	Processor unit NVIDIA P100 ▽	
Count 8 ▽	Cluster size 2	Interconnect InfiniBand FDR ▽

Software

Framework Caffe2 ▽	Batch size 1024 ▽	Scaling strong ▽

Add

Training performance

	Data	Hardware	Software	Time (hours)	
🟧	Size Epochs Model 10000000 50 AlexNet	Server PU Apollo 6500 NVIDIA P100 Count Cluster size Interconnect 8 2 IB	Framework Caffe2 Batch 1024(strong)	22.5	✕
🟪	Size Epochs Model 10000000 50 GoogleNet	Server PU Apollo 6500 NVIDIA P100 Count Cluster size Interconnect 8 2 IB	Framework Caffe2 Batch 1024(strong)	26.3	✕
🟦	Size Epochs Model 10000000 50 VGG19	Server PU Apollo 6500 NVIDIA P100 Count Cluster size Interconnect 8 2 IB	Framework Caffe2 Batch 1024(strong)	147.4	✕

Remove all

Thank you

Natalia Vassilieva
nvassilieva@hpe.com

Sergey Serebryakov
sergey.serebryakov@hpe.com

Sorin Cheran
sorin.cheran@hpe.com

Bruno Monnet
bruno.monnet@hpe.com