IBM Systems

Deep Learning Server Solution 1편

Caffe Demos

The Caffe neural network library makes implementing state-of-the-art computer vision systems easy.

Classification

Click for a Quick Example

Maximally accurate | Maximally specific

car	2.84969
motor vehicle	2.48797
self-propelled vehicle	2.20207
wheeled vehicle	1.88237
sports car	1.85357

CNN took 0.804 seconds.

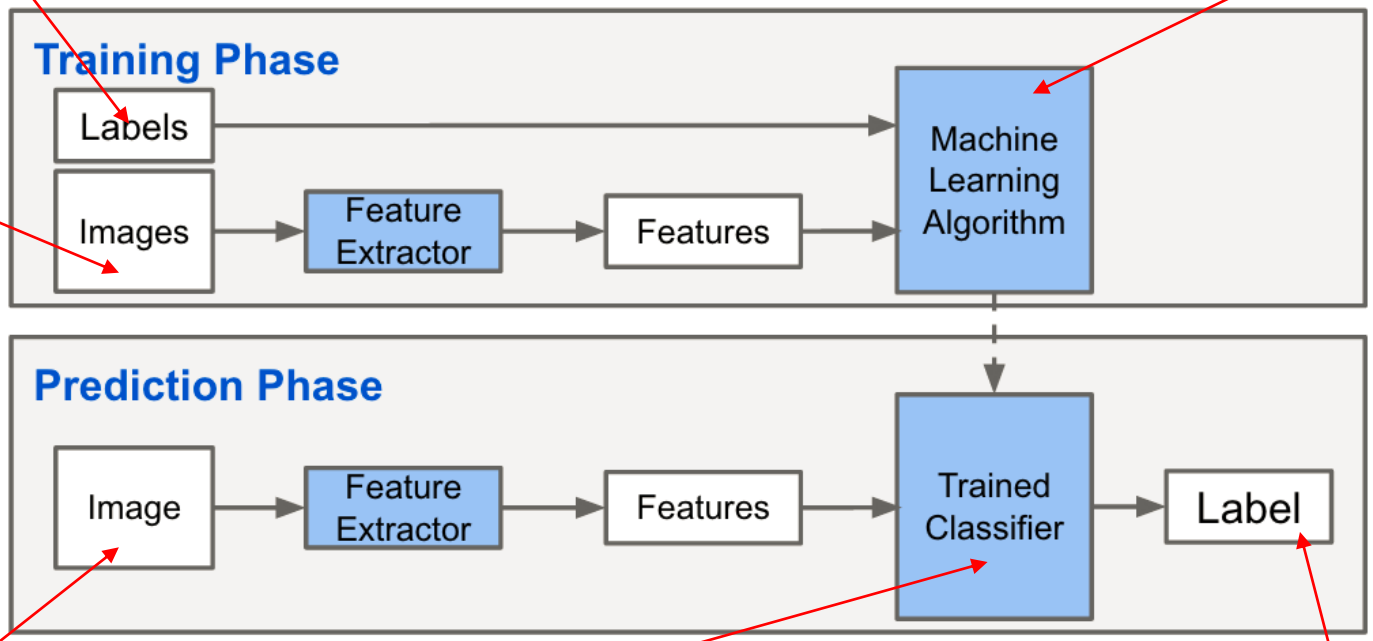
Provide an image URL

Classify URL

어떻게 caffe는 사물을 알아 보는가 ?

Caffe libraries
deploy.prototxt
train_val.prototxt
solver.prototxt

synset_words.txt
(1000 lines)

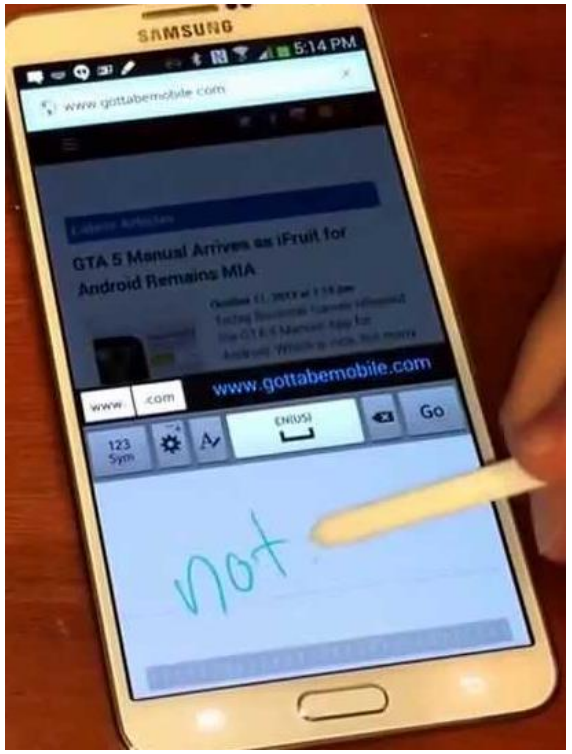


Caffe libraries
bvlc_reference_caffenet.caffemodel (size 243MB)

seabird
aquatic bird
pelican
pelecaniform seabird
bird

Deep Learning의 training 성능 차이 : MNIST 사례

- The MNIST DATABASE of handwritten digits yann.lecun.com/exdb/mnist/
- Training set 6만개, test set 1만개의 손글씨 이미지
- Caffe나 Tensorflow의 예제 모델로 포함된 대표적 deep learning network



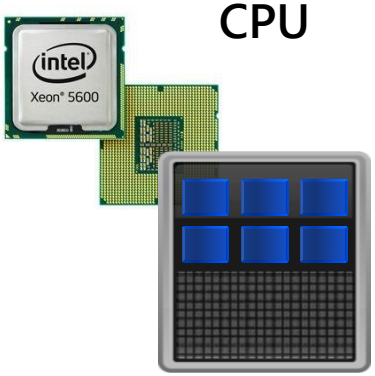

```
$ time ./examples/mnist/train_lenet.sh
...
I1116 01:46:51.764678 115864 solver.cpp:343] Iteration 10000, loss = 0.0051222
I1116 01:46:51.764695 115864 solver.cpp:363] Iteration 10000, Testing net (#0)
I1116 01:46:51.850540 115864 solver.cpp:442]     Test net output #0: accuracy =
0.9899
I1116 01:46:51.850567 115864 solver.cpp:442]     Test net output #1: loss = 0.0307451
(* 1 = 0.0307451 loss)
I1116 01:46:51.850576 115864 solver.cpp:348] Optimization Done.
I1116 01:46:51.850589 115864 caffe.cpp:291] Optimization Done.
```

Processor	CPU (POWER8)	Kepler GPU (K80)	PASCAL GPU (P100)
Training 시간	52m22.433s	0m43.963s	0m22.560s

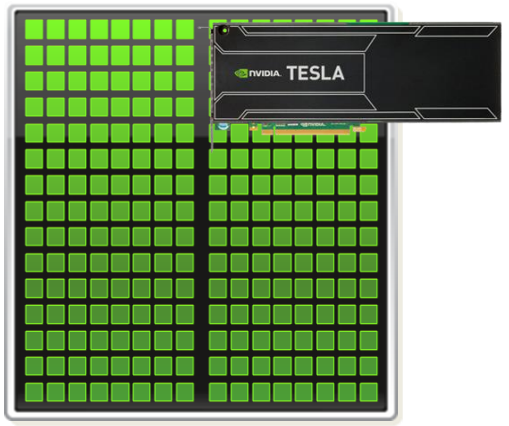
Source : IBM internal test

CPU와 GPU의 비교

CPU



GPU



Intel Xeon E5-2690 v3:

Clock speed: 2.6 GHz
12 cores

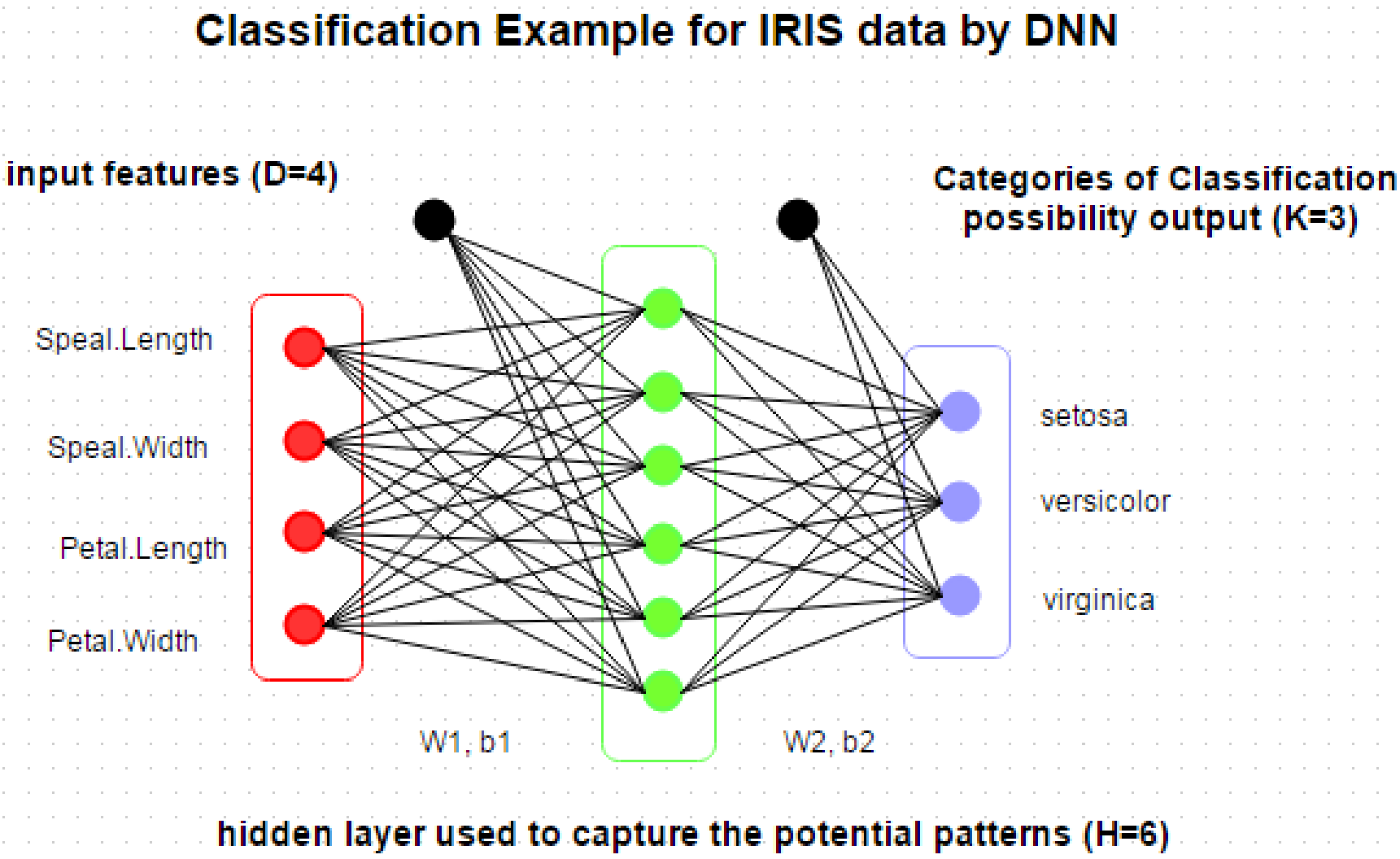
Up to 0.6 GFLOPS double precision
Memory size: 768 GB
Bandwidth: 68 GB/sec

NVIDIA Tesla K80:

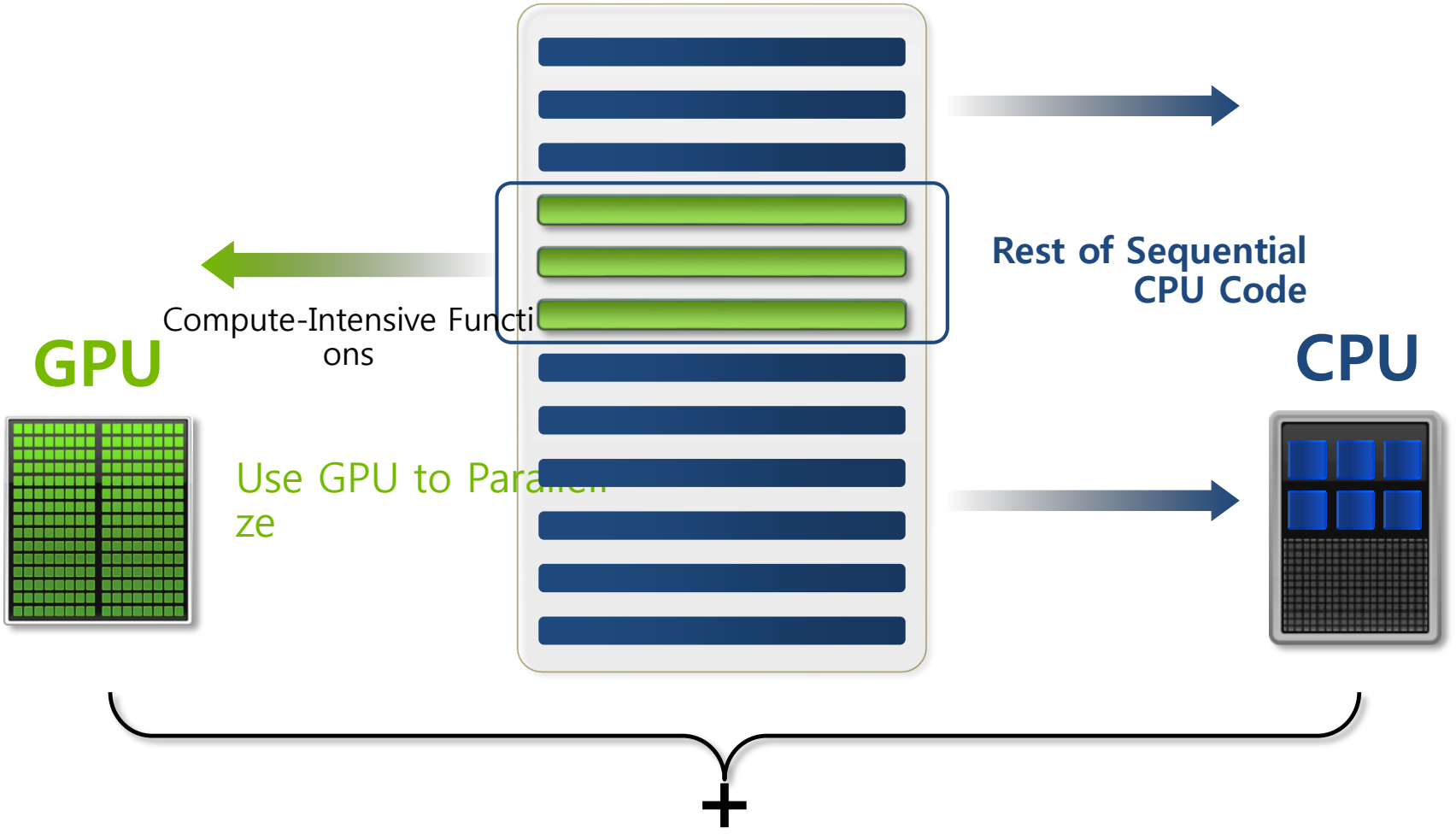
Clock speed: 560MHz
4992 CUDA cores (2496 per GPU)
8.73 TFLOPS single precision
2.91 TFLOPS double precision
Memory size: 24 GB (12GB per GPU)
Bandwidth: 480 GB/sec

IBM

왜 Deep Learning에 GPU가 적합한가 ?

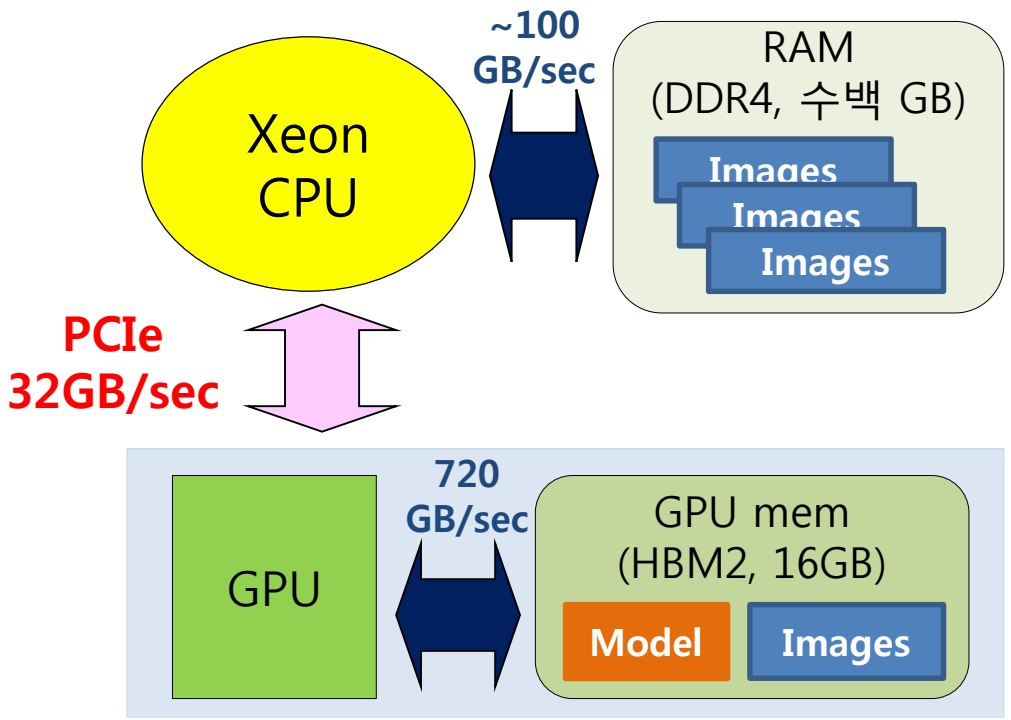


CPU와 GPU의 협업


Compute-Intensive Functi
ons
Rest of Sequential
CPU Code
GPU
CPU
Use GPU to Parallel
ze

GPU 컴퓨팅에서의 병목

GPU에서 하는 연산의 logic과 data는 모두 GPU memory 안에 일단 copy되어야 함



기존 GPU 컴퓨팅의 병목

- GPU board의 GPU 메모리 크기
  : 대개 12~24GB
  → Multi-GPU를 이용하여 해결
- GPU-CPU 간의 연결 속도
  : PCIe Gen3
  → 답이 없었으며, 해결을 위해서는 특별한 HW가 필요

CPU-GPU memcpy의 문제

(a, b, c, d, e, …. z) + (A, B, C, D, E, …. Z) = (a+A, b+B, c+C, d+D, e+E, …. z+Z)

calcCpu

```
for (int k = 0; k < tot; k++) {
c_B[k] = c_B[k] + c_A[k];
}
```

Vector computation time with CPU only
= calcCpu: **1662 ms**

calcGpu

```
cuMemcpyHtoD(d_A, Pointer.to(h_A), tot * Sizeof.DOUBLE);
add(tot, d_B, d_B, d_A);
cuMemcpyDtoH(Pointer.to(h_B), d_B, tot * Sizeof.DOUBLE);
```

Using GPU
= calcCpu: **174 ms**

copytoDevice: 1514
copytoHost: 1875
calcGpu: 174
Vector computation time with GPU = copytoDevice+copytoHost+calcGpu = **3563 ms**

POWER9 : Premier Acceleration Platform

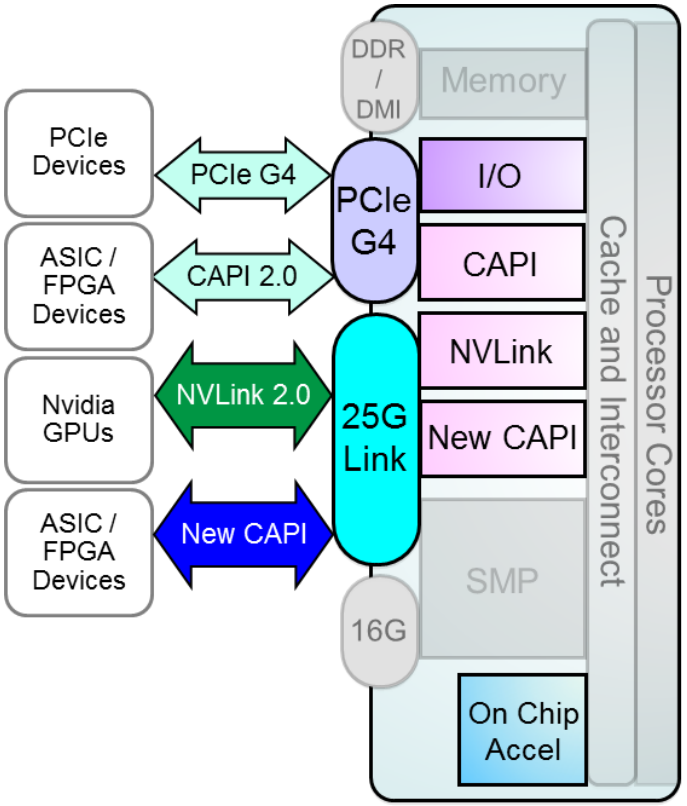
25G link에 의해 향상되는 NVLink 2.0과 OpenCAPI 3.0

최신 I/O 및 accelerator 연결 기술

- **PCIe Gen 4** x 48 lanes – 192 GB/s duplex
- **25G Link** x 48 lanes – 300 GB/s duplex

개방형 표준에 따른 견고한 가속 컴퓨팅 생태계

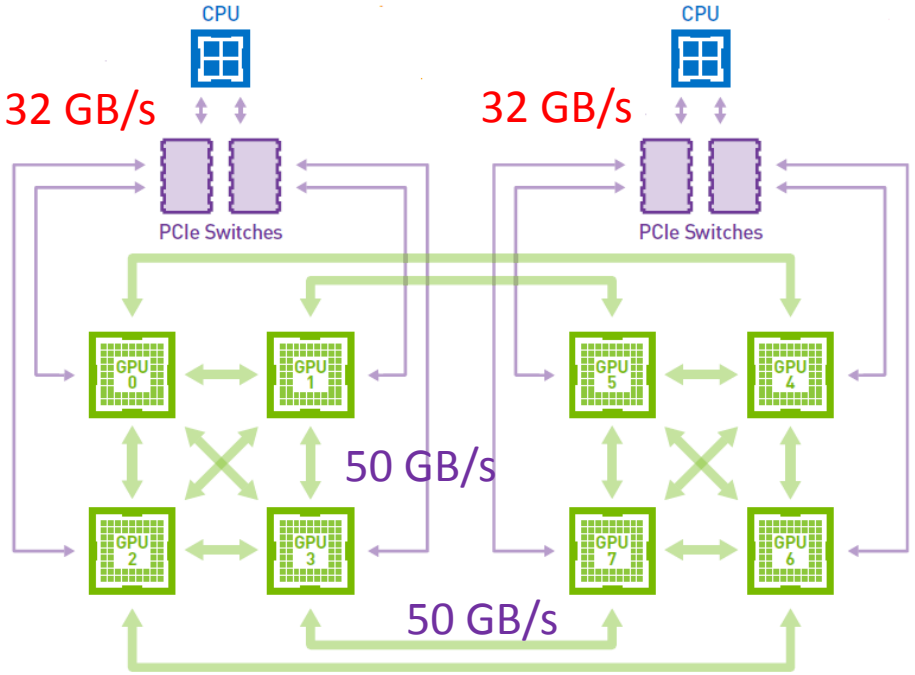
- **CAPI 2.0** – POWER8 대비 4배의 대역폭 (PCIe Gen4)
- **NVLink 2.0** – 차세대 GPU/CPU interconnect
  - NVLink1.0 대비 2배의 대역폭
  - 단순해지는 programming model
    - Coherency, virtual addressing, 낮은 overhead
- **OpenCAPI 3.0** – 높은 대역폭, 낮은 low latency, FPGA 등을 위한 개방형 interface



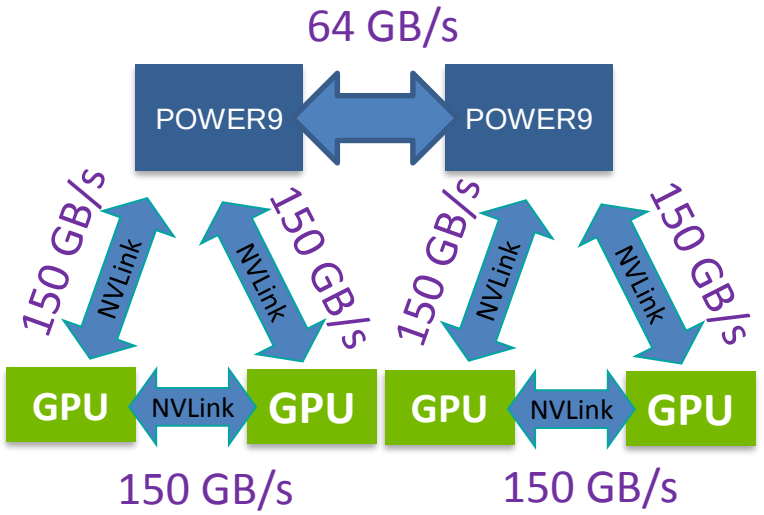
Source : https://openpowerfoundation.org/wp-content/uploads/2016/11/Jeff-Stuecheli-POWER9-chip-technology.pdf

NVLink 아키텍처의 비교

CPU-GPU 간의 NVLink, 그리고 NVLink *3 = 150 GB/sec가 Minsky의 특장점



x86 GPU



AC922

- CPU와 GPU간은 PCIe로 연결 (32GB/sec)
- 4개 GPU끼리 NVLink * 1 link로 연결 (50GB/sec)
- 다른 socket의 GPU 4개와의 연결은 2-hop 구조 (NVIDIA NCCL library에 의한 최적화 P2P)

- CPU와 GPU간을 NVLink * 3 link로 연결 (150GB/sec)
- 2개 GPU끼리 NVLink * 3 link로 연결 (150GB/sec)
- 다른 socket의 GPU 2개와의 연결은 64GB/s(4 byte * 16GHz)의 SMP X bus로 연결 (NVIDIA NCCL library에 의한 최적화 P2P)

IBM

CORAL project의 수퍼컴 노드 AC922 “Newell”

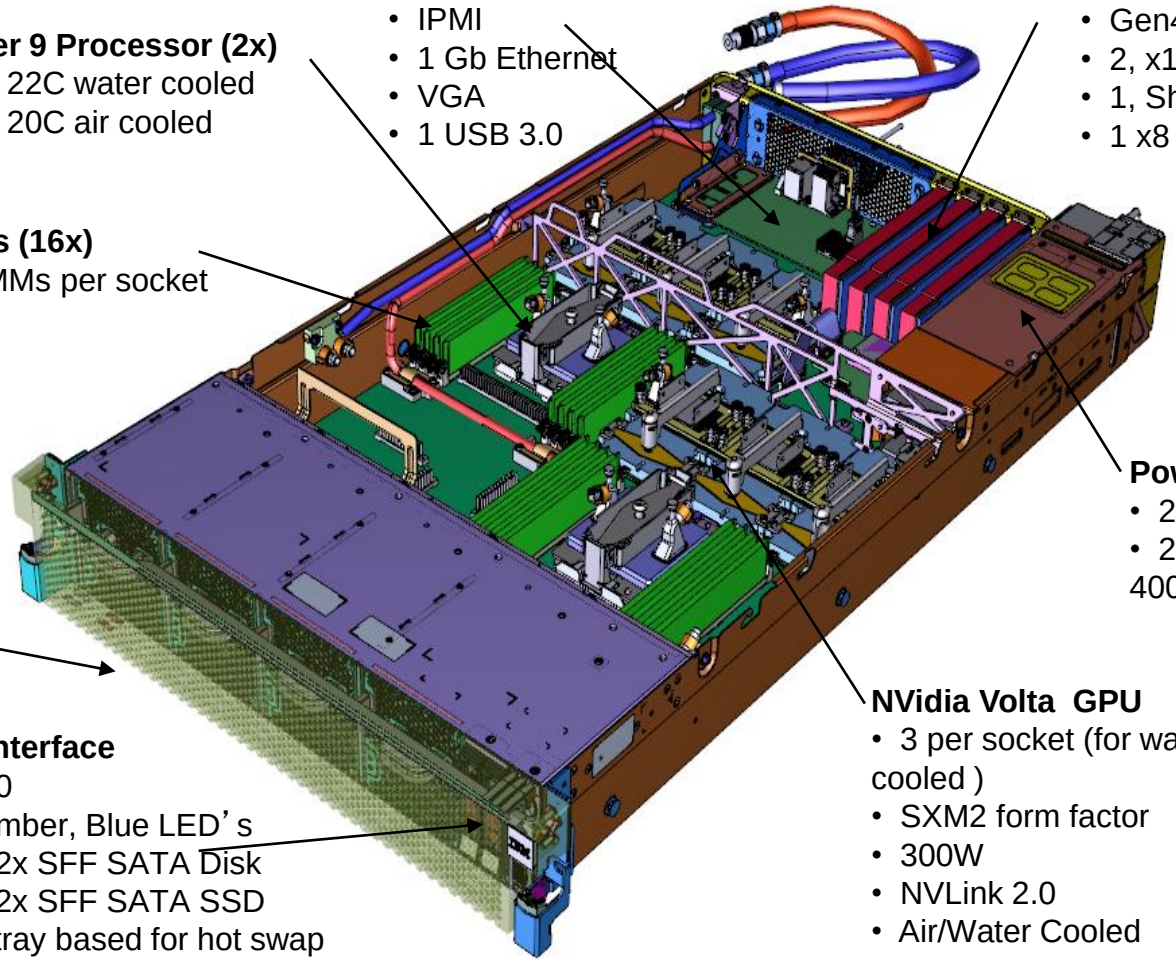
POWER9과 Volta를 NVLink 2.0을 통해 150GB/s로 연결

Power 9 Processor (2x)
- 18, 22C water cooled
- 16, 20C air cooled

BMC Card
- IPMI
- 1 Gb Ethernet
- VGA
- 1 USB 3.0

PCIe slot (4x)
- Gen4 PCIe
- 2, x16 HHHL Adapter
- 1, Shared slot
- 1 x8 HHHL Adapter

Memory DIMM’s (16x)
- 8 DDR4 IS DIMMs per socket



Power Supplies (2x)
- 2200W
- 200VAC, 277VAC, 400VDC input

Cooling Fans
- Counter- Rotating
- Hot swap
- 80mm shown

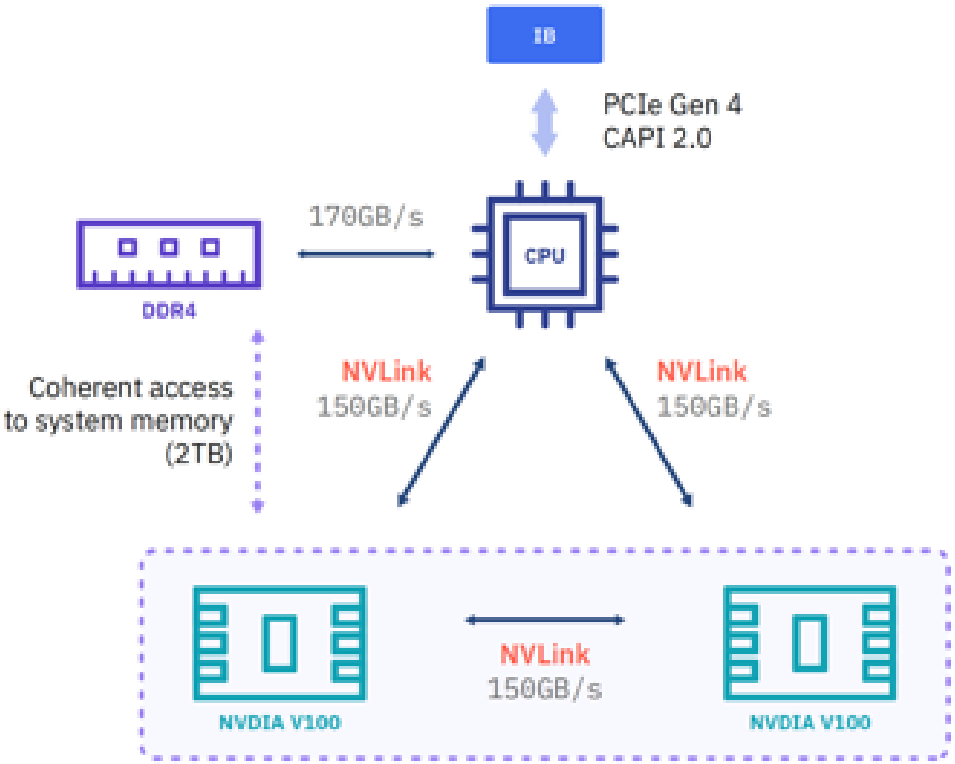
NVidia Volta GPU
- 3 per socket (for water-cooled)
- SXM2 form factor
- 300W
- NVLink 2.0
- Air/Water Cooled

Operator Interface
- 1 USB 3.0
- Green, Amber, Blue LED’s
- Optional 2x SFF SATA Disk
- Optional 2x SFF SATA SSD
- Disk are tray based for hot swap

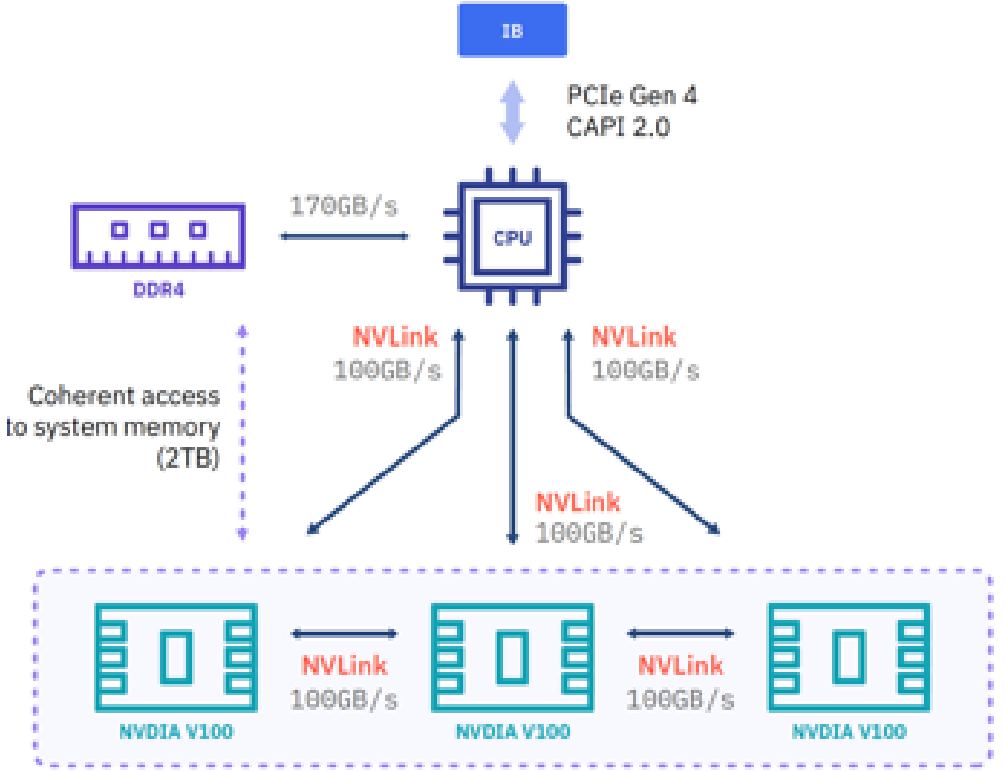
AC922 “Newell”의 2가지 옵션

4 GPU와 6 GPU의 2가지 구성안이 가능

4 GPUs – 공랭식
연결간 대역폭 150GB/s

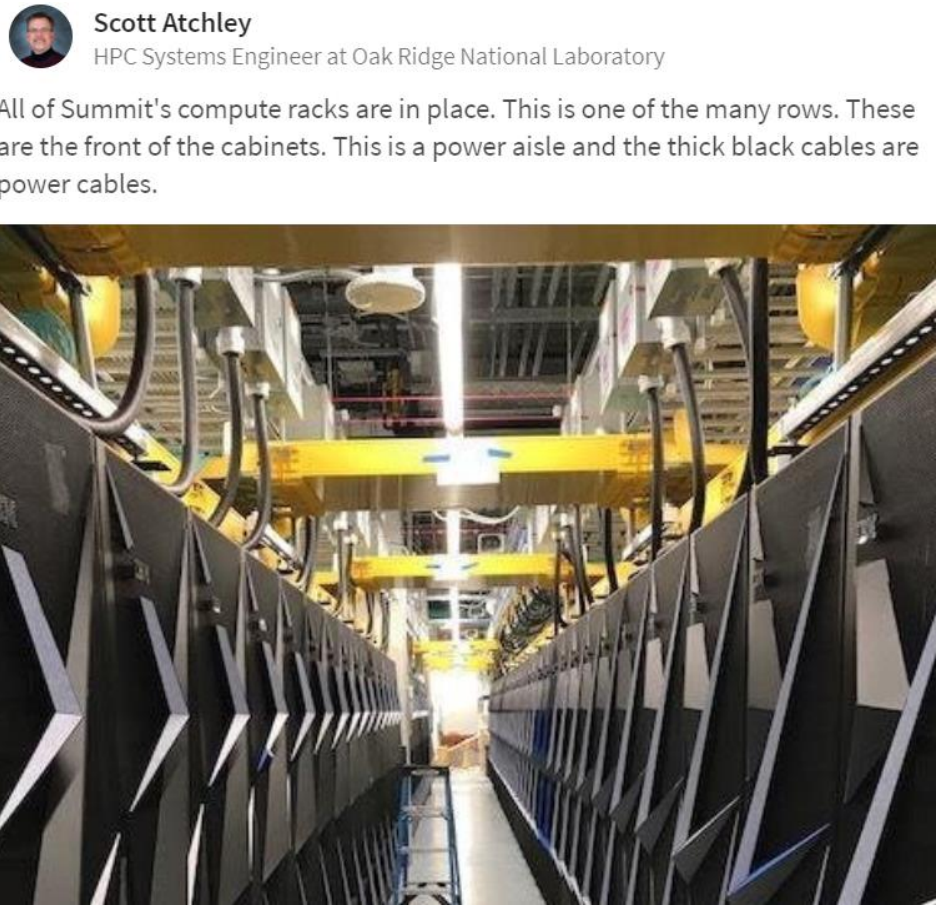


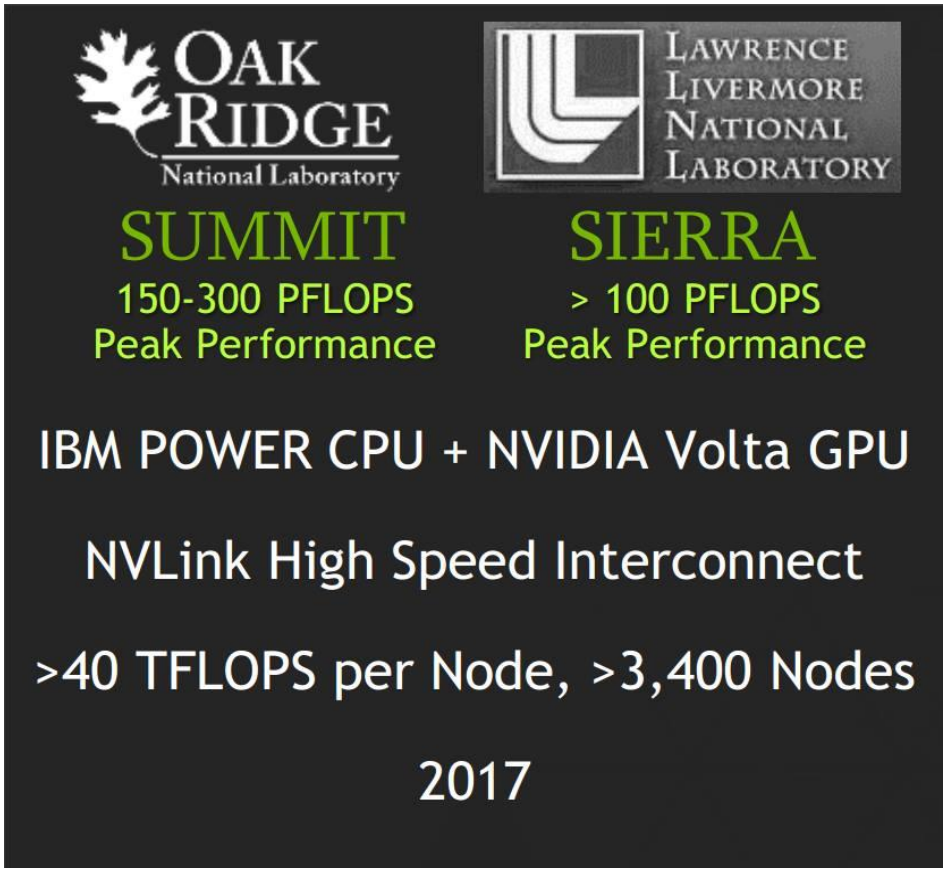
6 GPUs – 수냉식
연결간 대역폭 100GB/s



To-be #1 in TOP500 : CORAL project

POWER9 + Volta 기반의 수퍼컴 노드 “Newell”로 구성되는 Summit과 Sierra

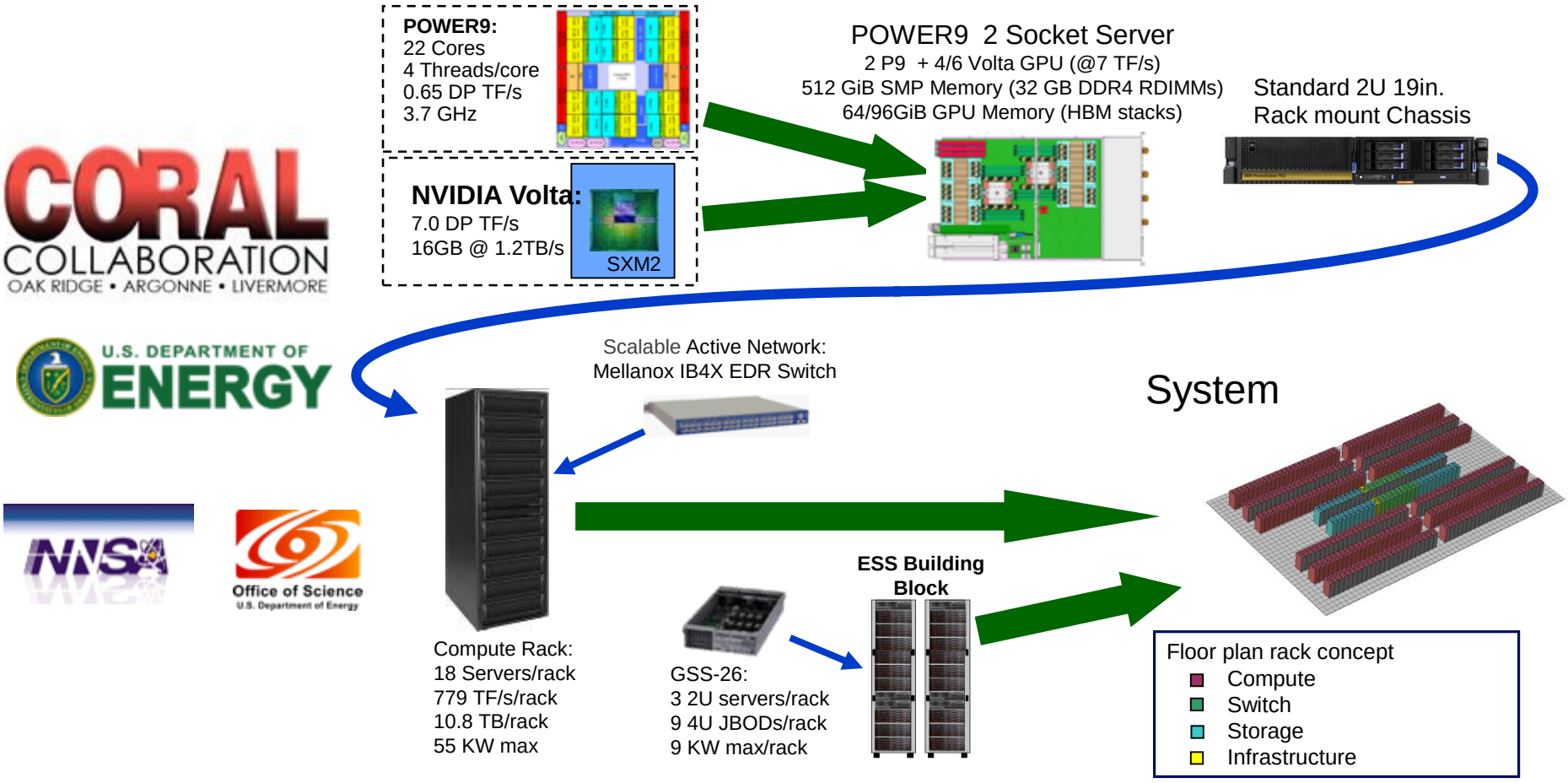




IBM

CORAL project의 수퍼컴 구조

POWER9 + Volta 기반의 AC922와 Spectrum Scale 병렬 파일시스템으로 구성



14

Minsky vs. AC922 성능 비교

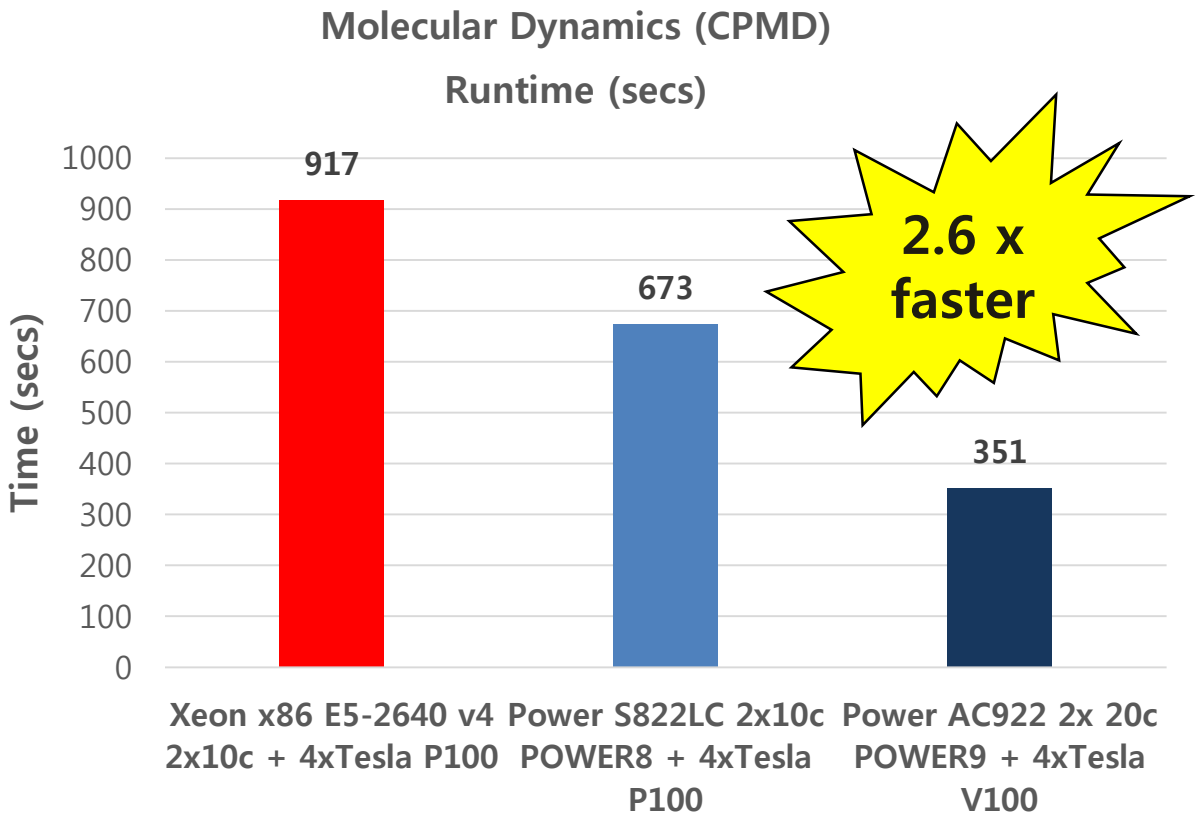
GPU FLOPS는 1.5배 향상 + NVLink는 2배 가까이 향상 = 몬테카를로 성능 4.5배 향상

CUDA samples	Measurement	Minsky P100 * 4	AC922 V100 * 4	AC922/Minsky ratio
simpleMultiCopy	Memcpy host to device (GB/s)	32.772	66.797	204%
	Memcpy device to host (GB/s)	33.415	63.581	190%
	Kernel (GB/s)	1338.152	2408.305	180%
	Fully serialized execution (GB/s)	29.483	56.879	193%
	Overlapped using 4 streams (GB/s)	60.872	121.408	199%
simpleP2P	cudaMemcpy between GPU0 and GPU1 (GB/s)	32.9	65.05	198%
MonteCarloMultiGPU	Total time (ms.)	46.811	10.364	22%
	Options per sec	700006.40	3161713.53	452%
BlackScholes	BlackScholesGPU() time (msec)	0.161	0.100	62%
	Effective memory bandwidth (GB/s)	497.027	797.974	161%
	Throughput (GOptions/s)	49.703	79.797	161%

AC922 “Newell”의 NVLink 2.0에 의한 획기적 성능 향상

기존 x86 기반 P100 GPU보다 2.6배, POWER8 기반 P100 GPU보다 1.9배의 성능 향상

- 분자역학 code인 CPMD는 TB 단위의 data가 CPU와 GPU 사이를 이동
  - 이로 인해 CPU-GPU 병목이 걸리는 대표적인 업무
  - PCIe에서는 3.3TB 이동에 300초 이상
  - NVLink 2.0에서는 70초
- P100과 V100의 이론상 성능 차이는 1.5배
- 실제 성능 차이는 2.6배

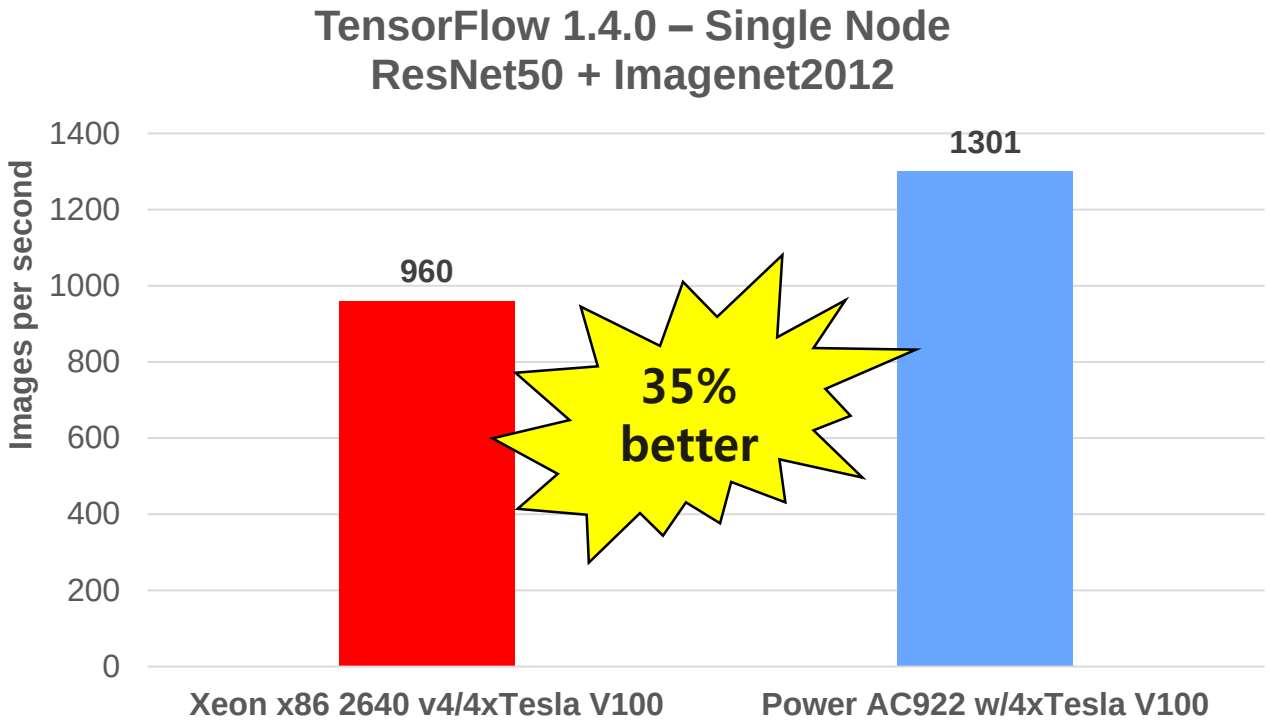


IBM

AC922 “Newell”의 Deep Learning 성능 비교

POWER9 + Volta + NVLink 2.0을 통해 차별화된 Deep Learning 성능

- **35% more images/sec .vs tested x86 systems**
- ResNet50 testing on ILSVRC 2012 dataset
  - Training on 1.2M images
  - Validation on 50K images

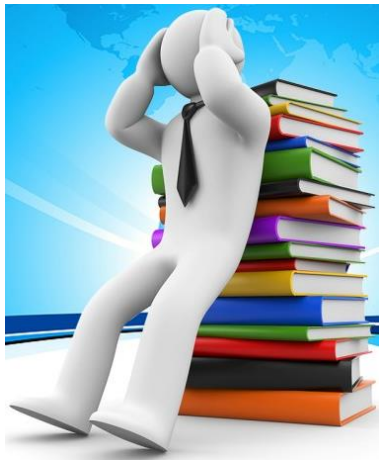


- Results are based IBM Internal Measurements running 1000 iterations of HPM Resnet50 on 1.2M images and validation on 50K images with Dataset from ILSVRC 2012 aka Imagenet 2012.
- Hardware: Power AC922; 40 cores (2 x 20c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU ; Red Hat Enterprise Linux 7.4 for Power Little Endian (POWER9). Competitive stack: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4x Tesla V100 GPU, Ubuntu 16.04.
- Software: Tensorflow 1.4 framework and HPM Resnet50. Found at mldl-repo-local-esp-5.0.0-5rc4.ppc64le.rpm and https://github.com/tensorflow/benchmarks.gif with the following parameters:Batch-Size: 64 per GPU ; Iterations: 1100; Data: synthetic and imagenet; local-parameter-device: gpu; variable-update: replicated

IBM

편리한 무료 딥 러닝 toolkit – IBM PowerAI

쉽고 안정적이고 성능 좋은 Deep Learning SW stack



OpenCV, hdf5, bazel, protobuf, lmdb 등등의 수 많은 기반 open source SW 를 일일이 build한 뒤 Caffe, Tensorflow 등을 설치

0.5~1일

기반 open source SW는 물론 Caffe, Tensorflow 등 주요 최신 framework을 최적화 build 된 무료 package로 제공

5~10분

The IBM PowerAI deep learning frameworks

Caffe

TensorFlow

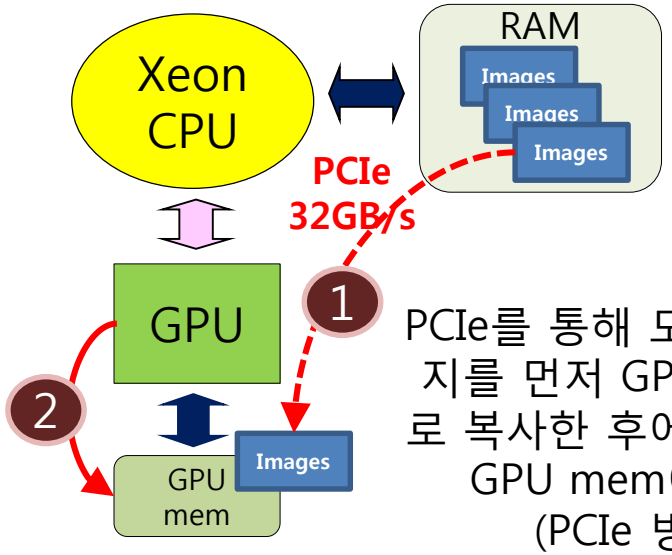
theano

torch

IBM Large Model Support (LMS) – host memory를 GPU에서 활용

GPU memory 크기의 한계 때문에 불가능하던 큰 모델의 training이 가능 !

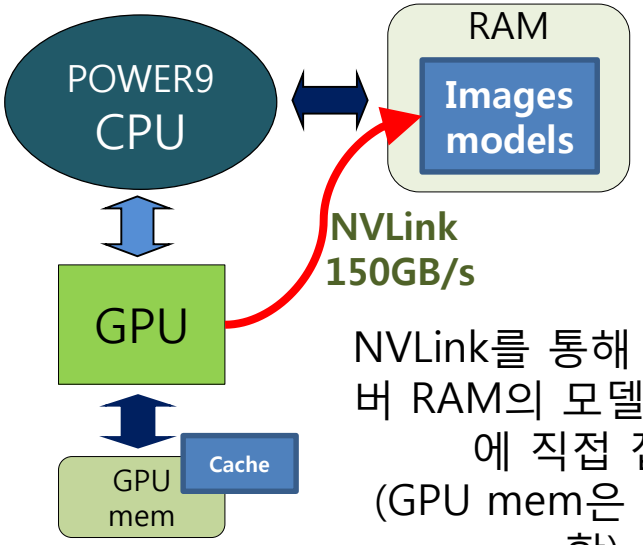
x86



PCIe를 통해 모델과 이미지를 먼저 GPU mem으로 복사한 후에야 GPU가 GPU mem에 접근 (PCIe 병목)

- GPU 메모리가 16GB로 제한되므로 작은 모델과 이미지만 사용 가능
- Batch size가 작아지므로 성능도 나빠짐
- 아예 training이 불가능한 경우도 발생

AC922



NVLink를 통해 GPU가 서버 RAM의 모델과 이미지에 직접 접근 (GPU mem은 cache 역할)

- 최대 1TB의 서버 RAM을 마치 GPU 메모리처럼 사용 가능하므로 훨씬 더 큰 모델과 이미지 사용 가능 (24배 더 큰 batch size도 가능)
- **CPU-GPU간 연결이 NVLink이기 때문**에 가능

IBM Large Model Support (LMS) – 더 큰 이미지도 처리 가능

"caffe time"에서 쉽게 확인되는 LMS의 편리함

```
$ caffe time -gpu 0 -model=bvlc_alexnet/deploy.prototxt --iterations=1
...
name: "AlexNet"
layer {
  name: "data"
  type: "Input"
  top: "data"
  input_param { shape: { dim: 10 dim: 3 dim: 1600 dim: 1200 } }
#  input_param { shape: { dim: 10 dim: 3 dim: 227 dim: 227 } }
}
```

error == cudaSuccess (2 vs. 0) out of memory

batch_size x channel x width x height

$ caffe time -gpu 0 -lms 16000000 -model=bvlc_alexnet/deploy.prototxt --iterations=1

...

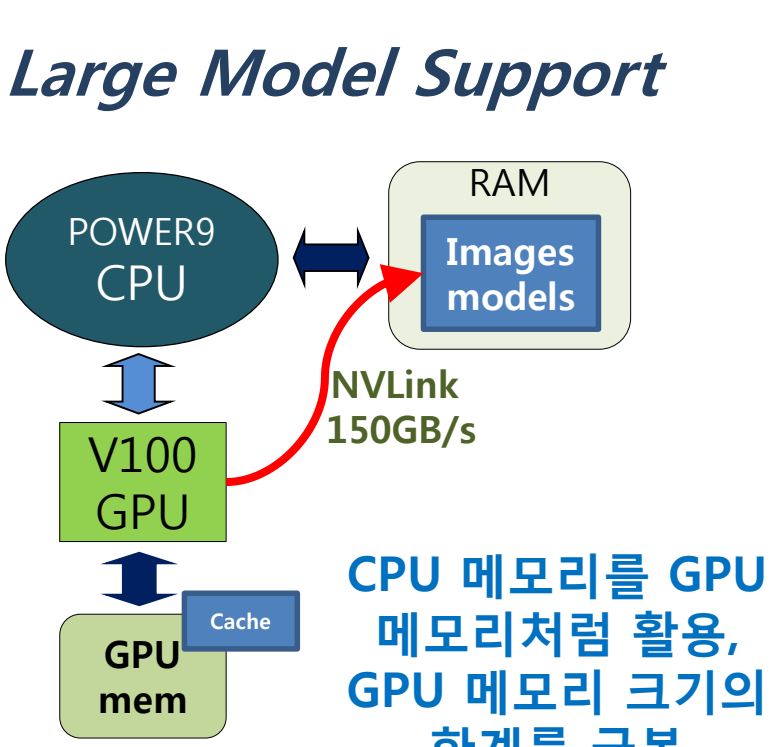
I1025 16:38:28.052325 29541 net.cpp:425] [LMS] conv4_forward [11] data:

...

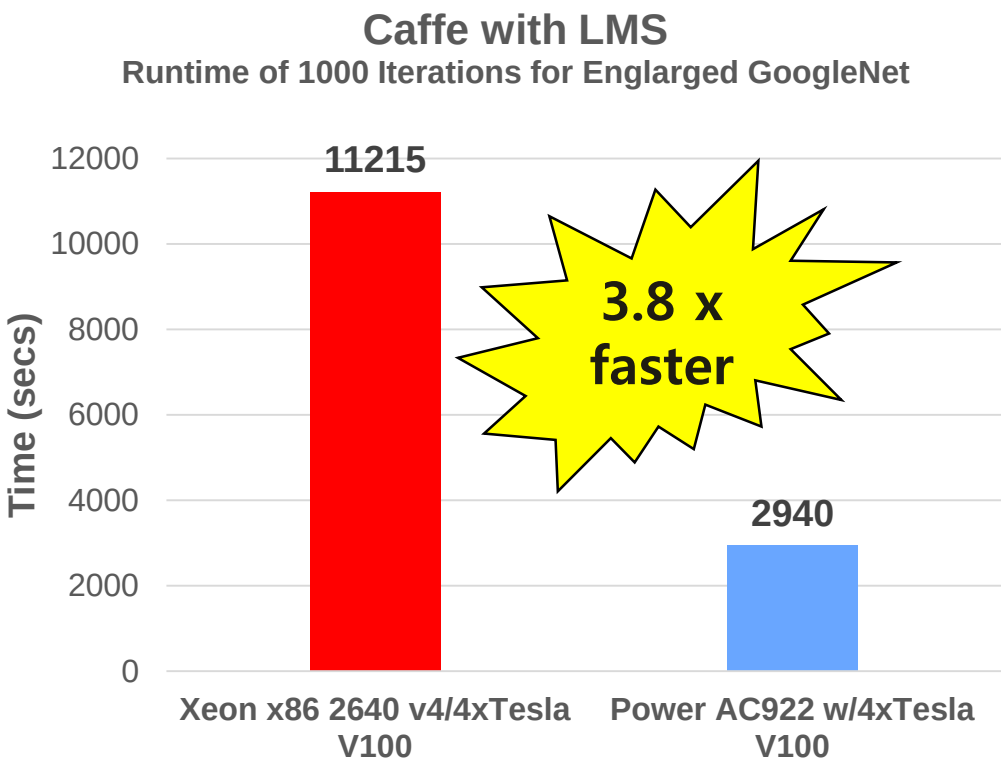
I1025 16:38:29.983943 29541 caffe.cpp:528] * Benchmark ends *****

AC922 “Newell”에서만 가능한 대형 이미지 처리

CPU-GPU 간의 NVLink 2.0을 이용한 PowerAI의 LMS (Large Model Support) 기능



CPU 메모리를 GPU 메모리처럼 활용, GPU 메모리 크기의 한계를 극복

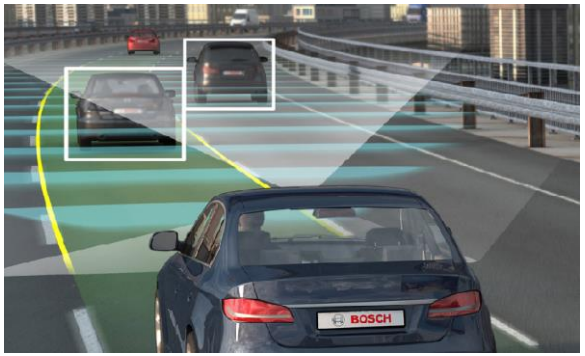


- Results are based IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240x2240) .
- Hardware: Power AC922; 40 cores (2 x 20c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU Pegas 1.0. Competitive stack: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04.
- Software: IBM Caffe with LMS Source code: https://github.ibm.com/TUNG/trlcaffe/tree/1.0-ibm-blc-bm-fix-hang+-p9collateral based on the branch "1.0-ibm-blc-bm-fix-hang+" (base for PowerAI R4) and a PR#5972 from BVLC/Caffe (for supporting cudnn7).

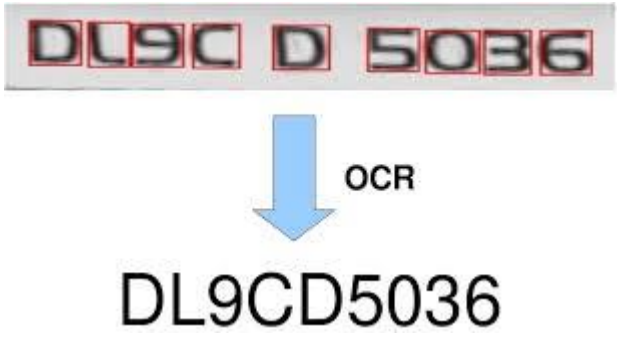
IBM

IBM GPU 솔루션 적용 사례

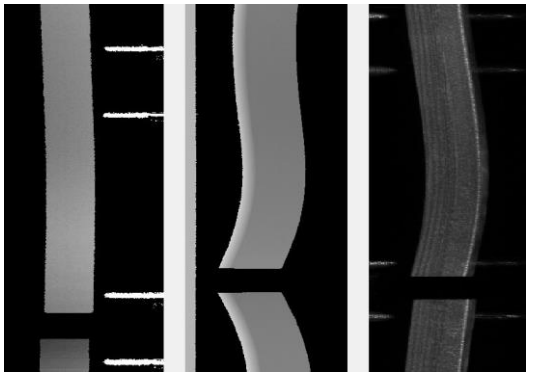
자동차 자율 주행



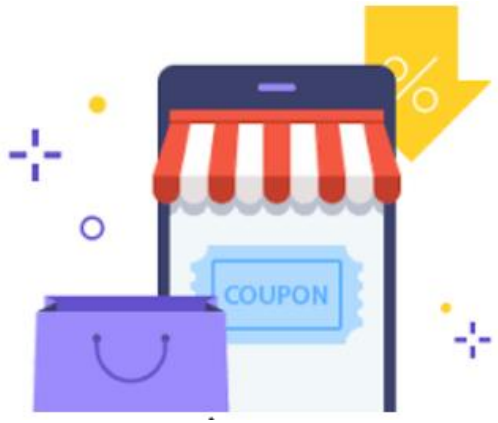
보험사 OCR



제조업 품질 검사

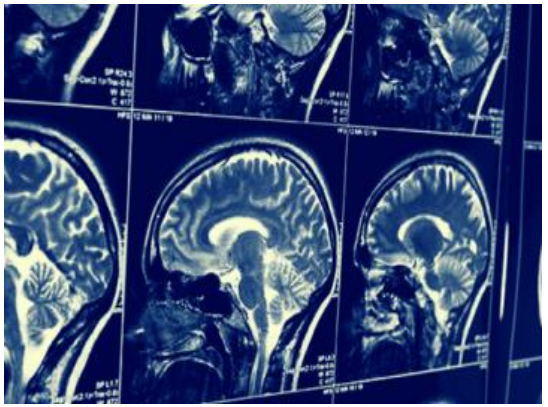


신용카드 고객 분석 및 오퍼링



금융사 Fraud Detection

Medical



Q & A