

“AI를 해주는 AI” H2O Driverless AI



Company Founded in Silicon Valley in 2012
Funded: \$75M. Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures

Products • H2O Open Source Machine Learning (14,000 organizations)
• H2O Driverless AI – Automatic Machine Learning

Leadership **Leader in Gartner MQ Machine Learning and Data Science Platform**

Team 100+ AI expertise (7 of the world's top 100 Kaggle Grandmasters/expert data scientists)

Global Mountain View, London, Prague, India

H₂O.ai



- ✓ 자동화:
 - Visualization
 - Feature Engineering
 - Model Tuning
 - Time Series
- ✓ 생성 모델의 편리한 활용 지원:
 - Automatic Pipelines
 - Low latency inferencing
- ✓ Machine Learning Interpretation:
 - 사유 부호(reason code) 지원
 - AI의 결정에 대한 해석과 설명이 가능
- ✓ Enterprise Ready:
 - 보안성 – LDAP, Kerberos
 - 확장성 – Scale with GPUs
 - 기업들의 source data를 지원



- ✓ 단순한 인터페이스
- ✓ Feature engineering을 자동화하여 정확성 증대
- ✓ 넓은 범위의 use case를 풀기 위해 자동화된 recipe들
- ✓ 적절한 model들의 집합을 찾고 조율하기 위한 자동화된 tuning

Driverless AI: top 5% in Amazon Kaggle competition

Driverless AI produces feature engineering pipeline (“more columns”) for downstream use



Amazon.com - Em...

Predict an employee's acc...
\$5,000 · 1,687 teams · 4...

Driverless AI: 8th place (out of 1687 - top 5%)

Driverless AI: Top-10 in BNP Paribas Kaggle competition



single run, **fully automated**: 2h on DGX Station! 6h on PC

BNP Paribas Cardif Claims Management

Can you accelerate BNP Paribas Cardif's claims management process?
\$30,000 · 2,926 teams · 2 years ago

Submission and Description

sub.csv
2 weeks ago by Ana Candel
94069f72/10/1 cv 0.4354 finished after 172 hrs

Private Score Public Score

0.42945 0.43156

Driverless AI: 10th place in private LB at Kaggle (out of 2926)

#	Rank	Team Name	Kernel	Team Members	Score (D)	Entries	Last
1	---	Dexter's Lab			0.42037	198	2y
2	---	escalated chi			0.43079	162	2y
3	---	Exploding Kitans			0.42182	124	2y
4	---	Brandon Nickel Utility			0.42239	251	2y
5	---	the flying burrito brothers			0.42450	264	2y
6	---	n_m			0.42335	4	2y
7	---	PAPY			0.42357	310	2y
8	---	KAME			0.42688	121	2y
9	---	Jack (Japan)			0.42744	22	2y
10	+1	Dmitry & Bohdan			0.43000	192	2y
11	+1	L1-Der			0.43005	95	2y
12	+2	BKIM2PDS			0.43089	338	2y
13	---	ix2xtd			0.43107	95	2y
14	---	Freemove			0.43148	134	2y
15	+1	Alma			0.43168	95	2y
16	+1	maze runners			0.43262	164	2y
17	---	BLR-2			0.43313	129	2y
18	+3	no one			0.43317	88	2y

2 months for Grandmasters — 2 hours for Driverless AI



- ✓ 규제 뿐만 아니라 디버깅을 위해 필요한 Interpretability
- ✓ 사유 부호(reason code)와 모델 interpretability를 영어 평문으로 생성
- ✓ 각 prediction에 대한 사유 부호 생성에 K-Lime, LOCO, partial dependence 등의 기술을 지원

Single Row Lookup DATASETS

Column: H2O Frame Row # Value: SEARCH Legend: Local Global

EXPORT TO CSV CLOSE

Summary

DAI Model ▼

Surrogate Models ▼

- KLIME
- Decision Tree
- Random Forest ▼

Dashboard

MLI Docs

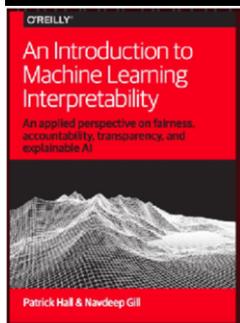
Scoring Pipeline

Global Reason Codes About reason codes ▼

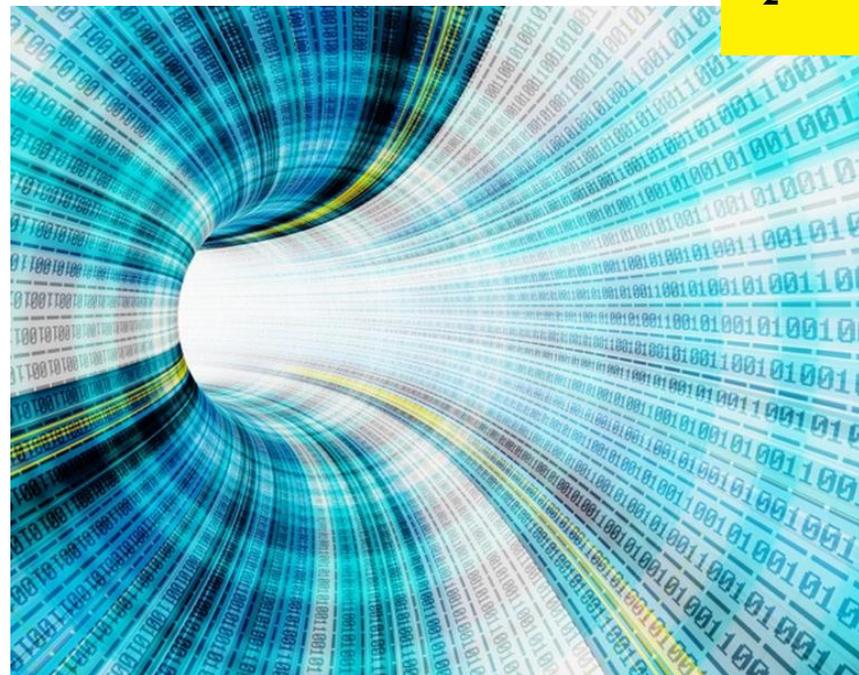
Global interpretable model explains 97.16% in predicted Hotels Occupancy rate (percent) for the entire dataset with RMSE = 1.347.

Variable	with value/ 1 unit increase (if blank)	is associated with predicted Hotels Occupancy rate (percent)	
Top Positive Global Attributions			
Total Occupancy rate (percent)		increase of	1.6
Motels Occupancy rate (percent)		increase of	0.15
Backpackers Occupancy rate (percent)		increase of	0.063
Negative Global Attributions			
Holiday Parks Occupancy rate (percent)		decrease of	1

Viewed 1 additional attributions, click to view all ...



- ✓ 독립된 prediction program의 자동 생성
 - Python 및 Java로 된 “scoring-pipeline” 자동 생성
 - 편리한 inferencing
- ✓ 새로운 model 생성시 편리한 update
- ✓ 복잡한 big data model에 대해 최적화된 scoring code
- ✓ 최말단 및 모바일 등 어떤 디바이스에서도 배치 가능한 간결한 scoring code
- ✓ 실시간 app을 만족시키는 millisecond 단위의 반응 속도

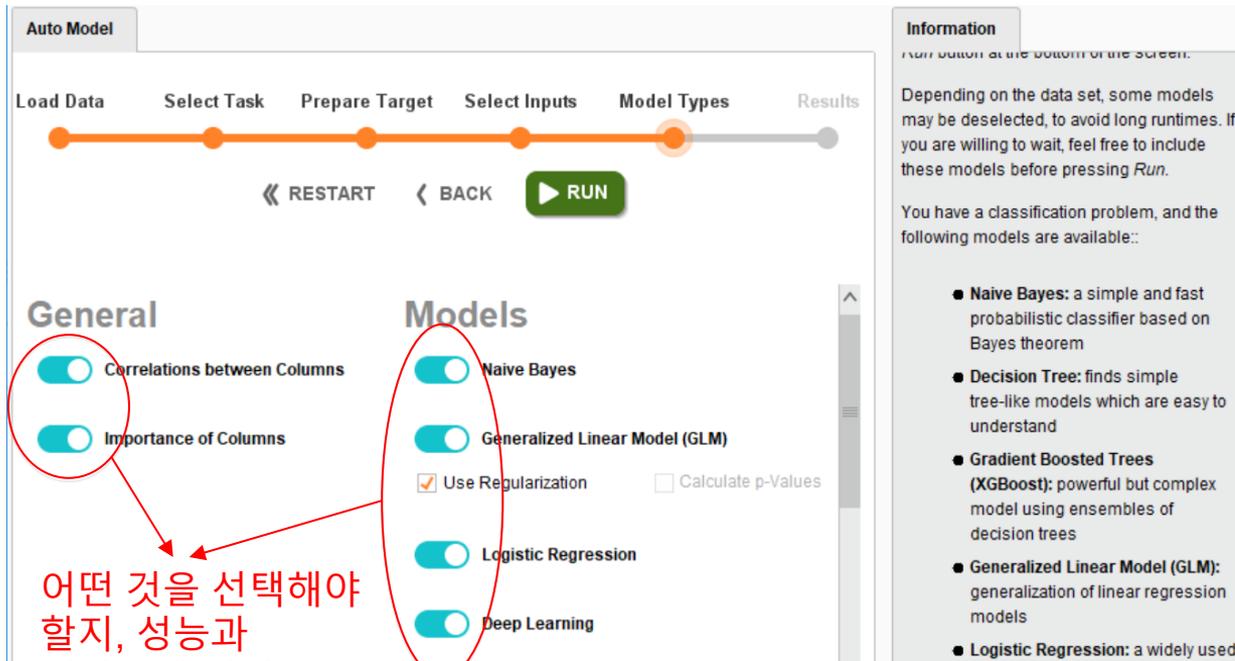


무엇을 택해야 할지 ?

무슨 값을 넣어야 할지 ?

- ✓ 모델 개발의 난제
 - Neural network 종류 선택
 - 각 flow variable 값의 설정
 - 기타 다수
- ✓ Try & error의 반복
 - 최적의 결과를 위해서는 반복이 필수
- ✓ 결국 Data scientist와 긴 시간이 필요
- ✓ **H2O는 최적의 feature engineering을 자동화로 쉽고 빠르게 해결**

Source : KNIME



General

- Correlations between Columns
- Importance of Columns

Models

- Naive Bayes
- Generalized Linear Model (GLM)
- Use Regularization Calculate p-Values
- Logistic Regression
- Deep Learning

Information

Depending on the data set, some models may be deselected, to avoid long runtimes. If you are willing to wait, feel free to include these models before pressing *Run*.

You have a classification problem, and the following models are available::

- **Naive Bayes:** a simple and fast probabilistic classifier based on Bayes theorem
- **Decision Tree:** finds simple tree-like models which are easy to understand
- **Gradient Boosted Trees (XGBoost):** powerful but complex model using ensembles of decision trees
- **Generalized Linear Model (GLM):** generalization of linear regression models
- **Logistic Regression:** a widely used

어떤 것을 선택해야 할지, 성능과 정확도에 어떤 영향을 줄지 ?

- ✓ 알고리즘의 선택
 - Dataset의 종류와 target에 따라 천차만별
 - 선택 가능한 recipe의 다양성도 중요
 - Accuracy와 run time에서 큰 차이의 결과
- ✓ 최적의 선택을 자동으로 해주는지 여부가 중요
 - Auto Model을 제공하는 제품에서도 주요 결정은 사용자의 몫
- ✓ H2O DAI는 수천개의 recipe 중에 주어진 과제에 최적의 것을 자동으로 파악하여 모델을 생성

Source : RapidMiner



Environment & Tools

- Node2Vec – Node representation learning
- Driverless AI – Feature Engineering & Model Training
- Spark – Data Preparation/Pre-processing
- Hardware – GPU server
 - 4 Pascal 100 GPU
 - 160 cores CPUs
 - 1 TB RAM

 © 2017 PayPal Inc. Confidential and proprietary.

17

H2O World 2017

H₂O.ai



Source : <https://youtu.be/r9S3xchrzIY>

Why H2O Driverless AI on IBM AC922 ?

2.6x

More RAM

Big Data Scale

9.5x

Max I/O bandwidth

High Speed Data Transfer

30x

Faster on GPUs

GPU Accelerated ML

2x

Data Ingest

Big Data Scale

1.5x

Feature Engineering

High Speed Data Transfer

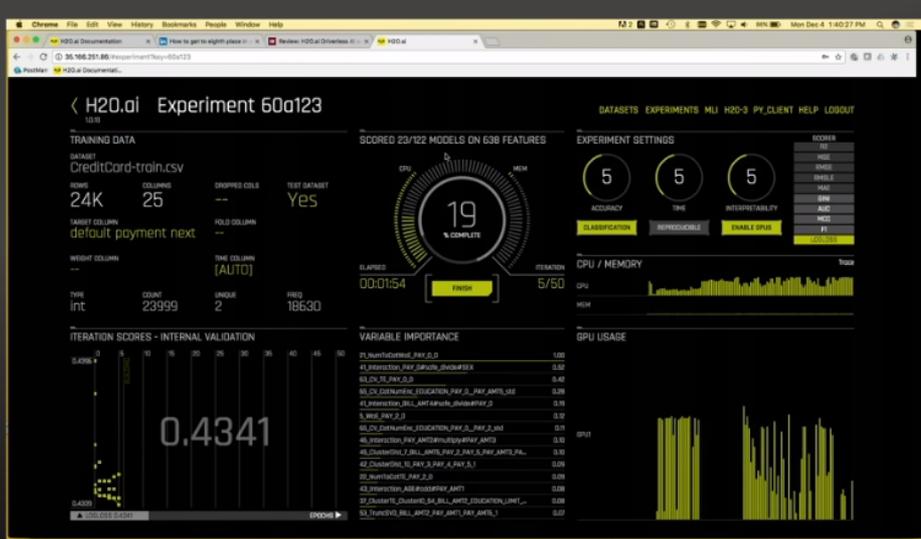
5x

Time Series

GPU Accelerated ML

NVLink와 PCIe Gen4를 탑재한 POWER9 프로세서

왜 H2O DAI에 POWER9 + V100 이 필요한가 ?



The screenshot shows the H2O.ai web interface for Experiment 60a123. The main dashboard displays training progress with a circular gauge at 19% complete. Key metrics include 24K rows, 25 columns, and a test dataset of 'Yes'. The interface also shows variable importance, GPU usage, and iteration scores. A red box highlights the text: "ship is actually faster than a GPU yes that's because it has to go over the PCI".

H2O World 2017

ship is actually faster than a GPU yes that's because it has to go over the PCI

“어떤 분들은 CPU가 GPU보다 훨씬 더 빠르다고 말씀하십니다. 예, 그건 2만줄 정도의 data를 **PCI Express 버스** 위에서 CPU와 GPU가 주고받아야 하기 때문이지요.”

Arno Candell, CTO,
H2O.ai
H2O World 2017

Source : <https://youtu.be/niiibeHJtRo>

H2O DAI의 병목은 GPU 성능이나 GPU 메모리가 아니라 연결 대역폭

Fri Oct 5 03:36:45 2018

```

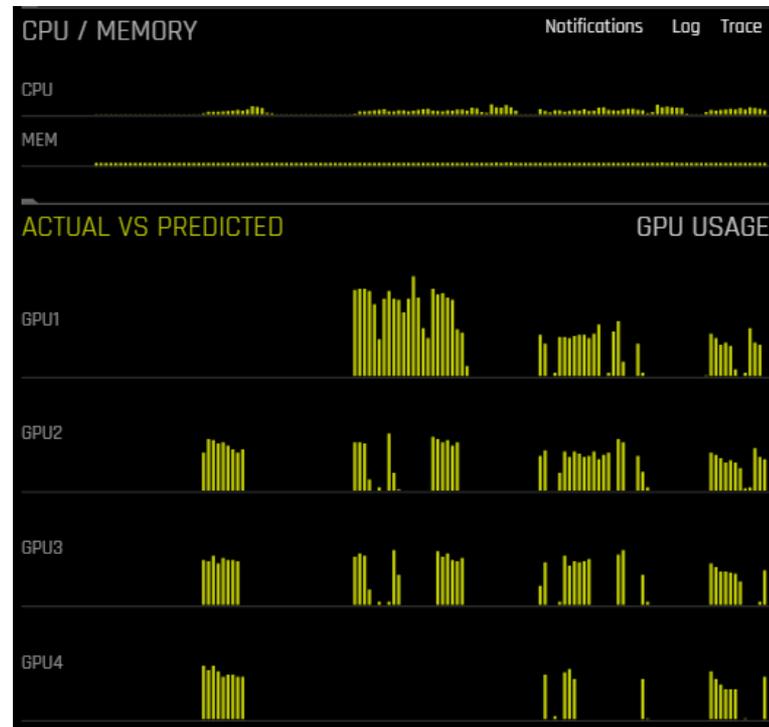
+-----+
| NVIDIA-SMI 396.26      Driver Version: 396.26      |
+-----+-----+
| GPU  Name          Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf   Pwr:Usage/Cap|  Memory-Usage | GPU-Util  Compute M. |
+-----+-----+-----+-----+-----+-----+
|  0  Tesla P100-SXM2...  On   | 00000002:01:00:0  Off |      0 |               |
|  N/A   34C    P0      64W / 300W |  455MiB / 16280MiB |    40%    Default |
+-----+-----+-----+-----+-----+-----+
|  1  Tesla P100-SXM2...  On   | 00000003:01:00:0  Off |      0 |               |
|  N/A   36C    P0      77W / 300W |  455MiB / 16280MiB |    38%    Default |
+-----+-----+-----+-----+-----+-----+
|  2  Tesla P100-SXM2...  On   | 0000000A:01:00:0  Off |      0 |               |
|  N/A   32C    P0      71W / 300W |  455MiB / 16280MiB |    40%    Default |
+-----+-----+-----+-----+-----+-----+
|  3  Tesla P100-SXM2...  On   | 0000000B:01:00:0  Off |      0 |               |
|  N/A   36C    P0      64W / 300W |  455MiB / 16280MiB |    38%    Default |
+-----+-----+-----+-----+-----+-----+

```

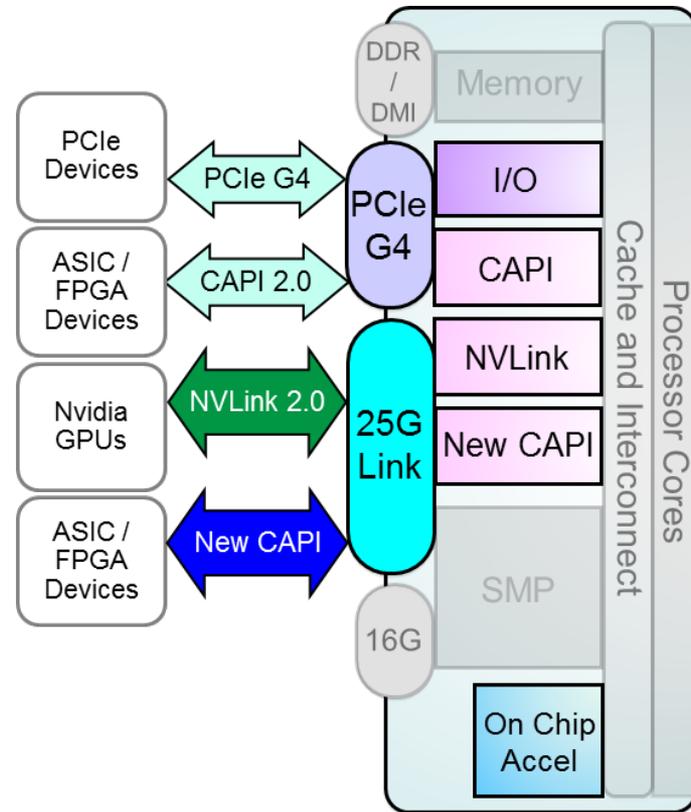
```

+-----+
| Processes:                      GPU Memory |
| GPU   PID  Type  Process name                        Usage  |
+-----+-----+-----+-----+-----+
|  0     94004  C    ...el-running (prot=False)-XGBoostModel-fit  445MiB |
|  1     94011  C    ...el-running (prot=False)-XGBoostModel-fit  445MiB |
|  2     94044  C    ...el-running (prot=False)-XGBoostModel-fit  445MiB |
|  3     94126  C    ...el-running (prot=False)-XGBoostModel-fit  445MiB |
+-----+-----+-----+-----+-----+

```

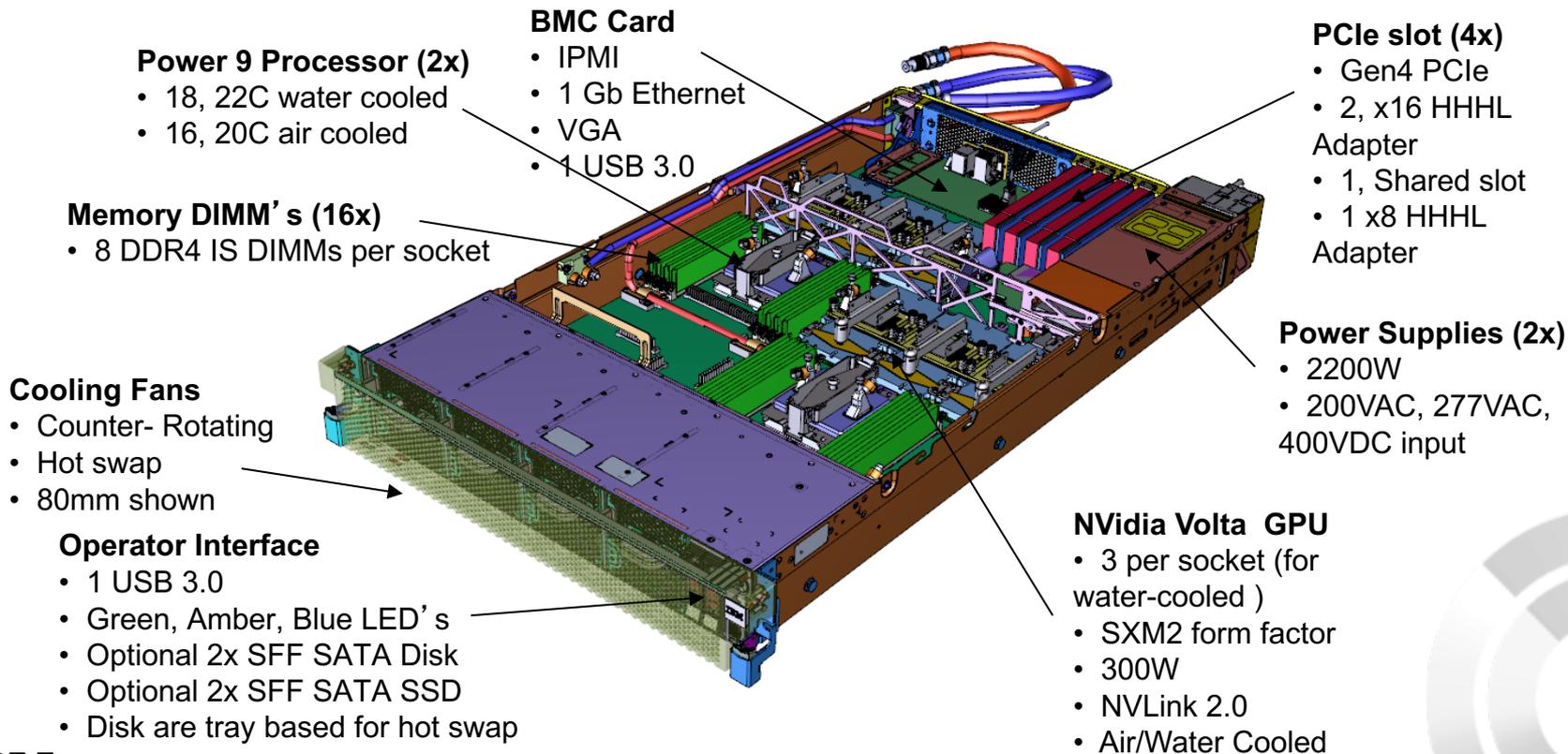


- ✓ 최신 I/O 및 accelerator 연결 기술
 - PCIe Gen 4 x 48 lanes – 192 GB/s duplex
 - 25G Link x 48 lanes – 300 GB/s duplex
- ✓ 개방형 표준에 따른 견고한 가속 컴퓨팅 생태계
 - CAPI 2.0 – POWER8 대비 4배의 대역폭 (PCIe Gen4)
 - NVLink 2.0 – 차세대 GPU/CPU interconnect
 - NVLink1.0 대비 2배의 대역폭
 - 단순해지는 programming model
 - OpenCAPI 3.0 – 높은 대역폭, 낮은 low latency, FPGA 등을 위한 개방형 interface

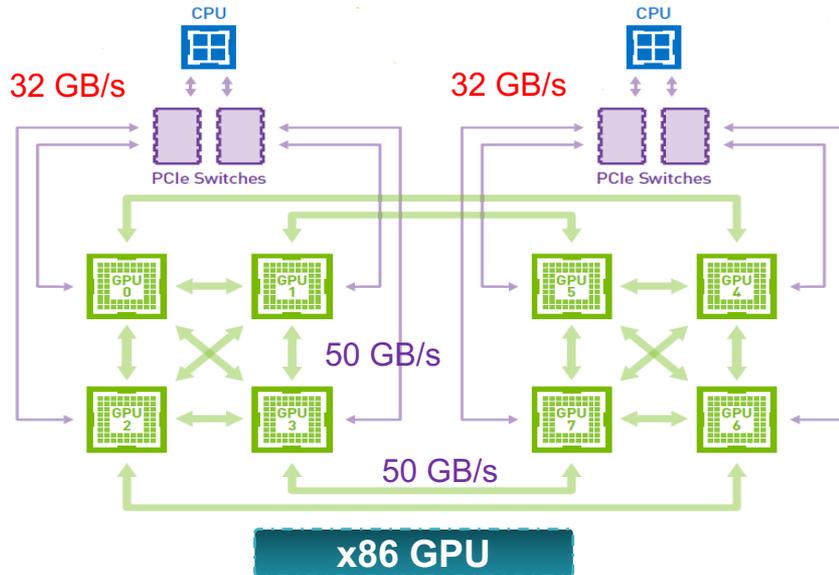


Source : <https://openpowerfoundation.org/wp-content/uploads/2016/11/Jeff-Stuecheli-POWER9-chip-technology.pdf>

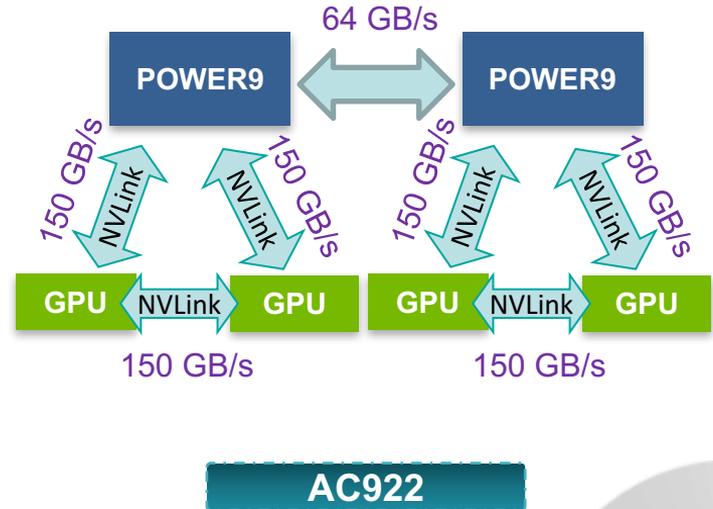
POWER9과 Volta를 NVLink 2.0을 통해 150GB/s로 연결



CPU-GPU 간의 NVLink, 그리고 NVLink * 3 = 150 GB/sec가 AC922의 특징점



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



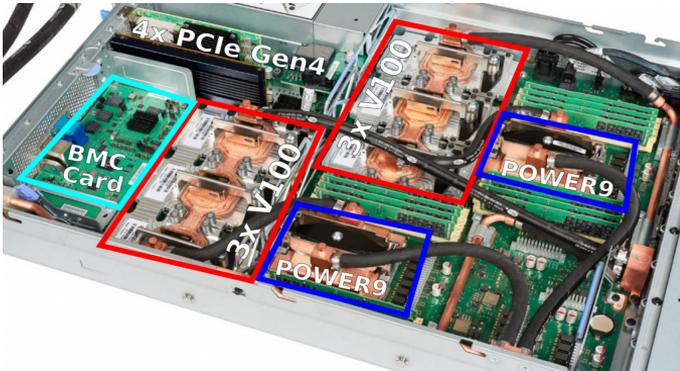
- CPU와 GPU간을 NVLink * 3 link로 연결 (150GB/sec)
- 2개 GPU끼리 NVLink * 3 link로 연결 (150GB/sec)
- 다른 socket의 GPU 2개와의 연결은 64GB/s(4 byte * 16GHz)의 SMP X bus로 연결

POWER9 + Volta 기반의 POWER9 서버로 구성되는 슈퍼컴 Summit과 Sierra



The US again has the world's most powerful supercomputer

The US just grabbed back the crown from China with the AI-focused Summit.



Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	187,659.3	8,806
2	Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
3	Sierra - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/NNSA/LLNL United States	1,572,480	71,610.0	119,193.6	