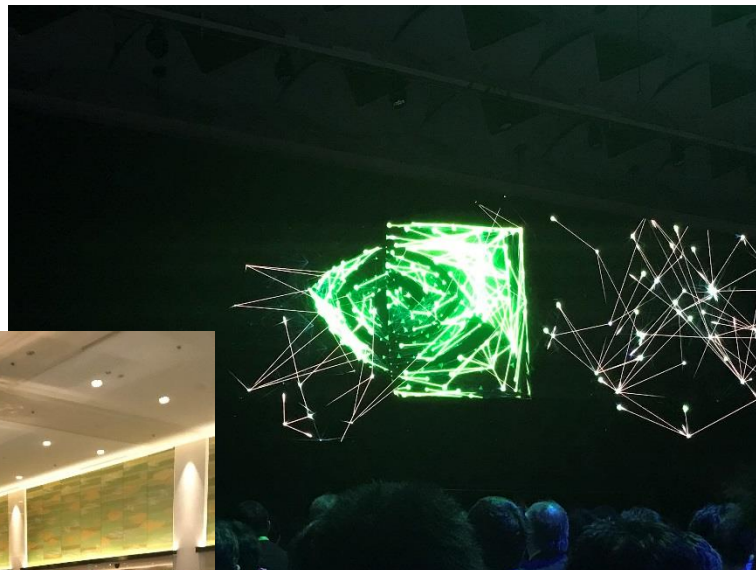


# GTC Japan 2018

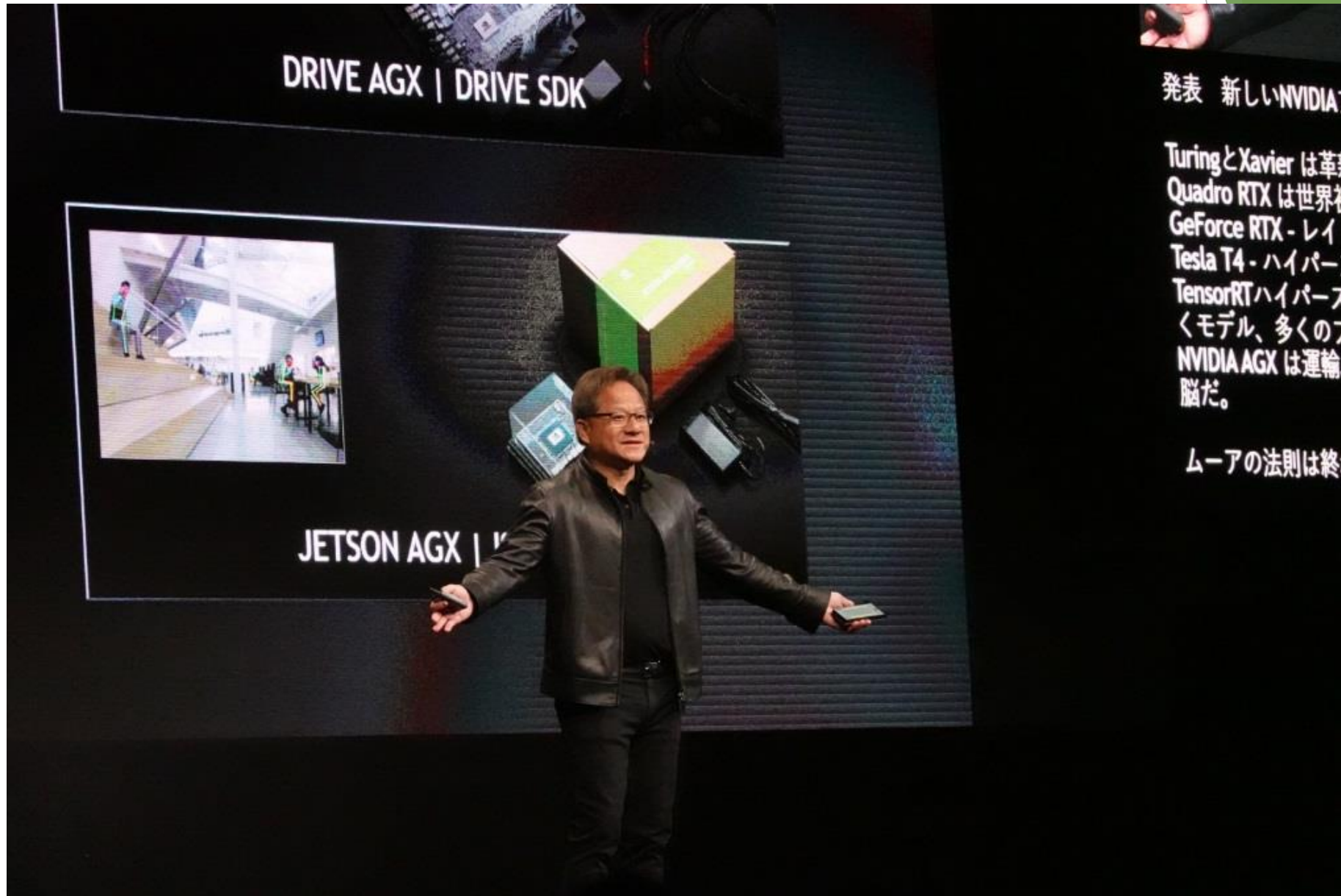
2018.09.13

# 会場：グランドプリンスホテル新高輪





# 基調講演：Jen-Hsun Huang (CEO)



# CG業界の夢であったリアルタイムレイトレーシングを実現するRTXの紹介

## NEW GEFORCE RTX GRAPHICS REINVENTED

### GEFORCE RTX 2070

8 TFLOPS | 8 TIOPS | 63 Tensor TFLOPS | 6 GIGA RAYS

FROM \$499

### GEFORCE RTX 2080

11 TFLOPS | 11 TIOPS | 85 Tensor TFLOPS | 8 GIGA RAYS

FROM \$699

### GEFORCE RTX 2080 Ti

14 TFLOPS | 14 TIOPS | 114 Tensor TFLOPS | 10 GIGA RAYS

FROM \$999



**NEW**  
**NVIDIA DGX-2**  
**THE LARGEST GPU EVER CREATED**

**2 PFLOPS**

**512GB HBM2**

**16 TB/sec Memory Bandwidth**

**10 kW | 160 kg**



# 日本のAI関連会社への納入について配慮した資料

LEADERS IN AI & HPC

Satellite Vision

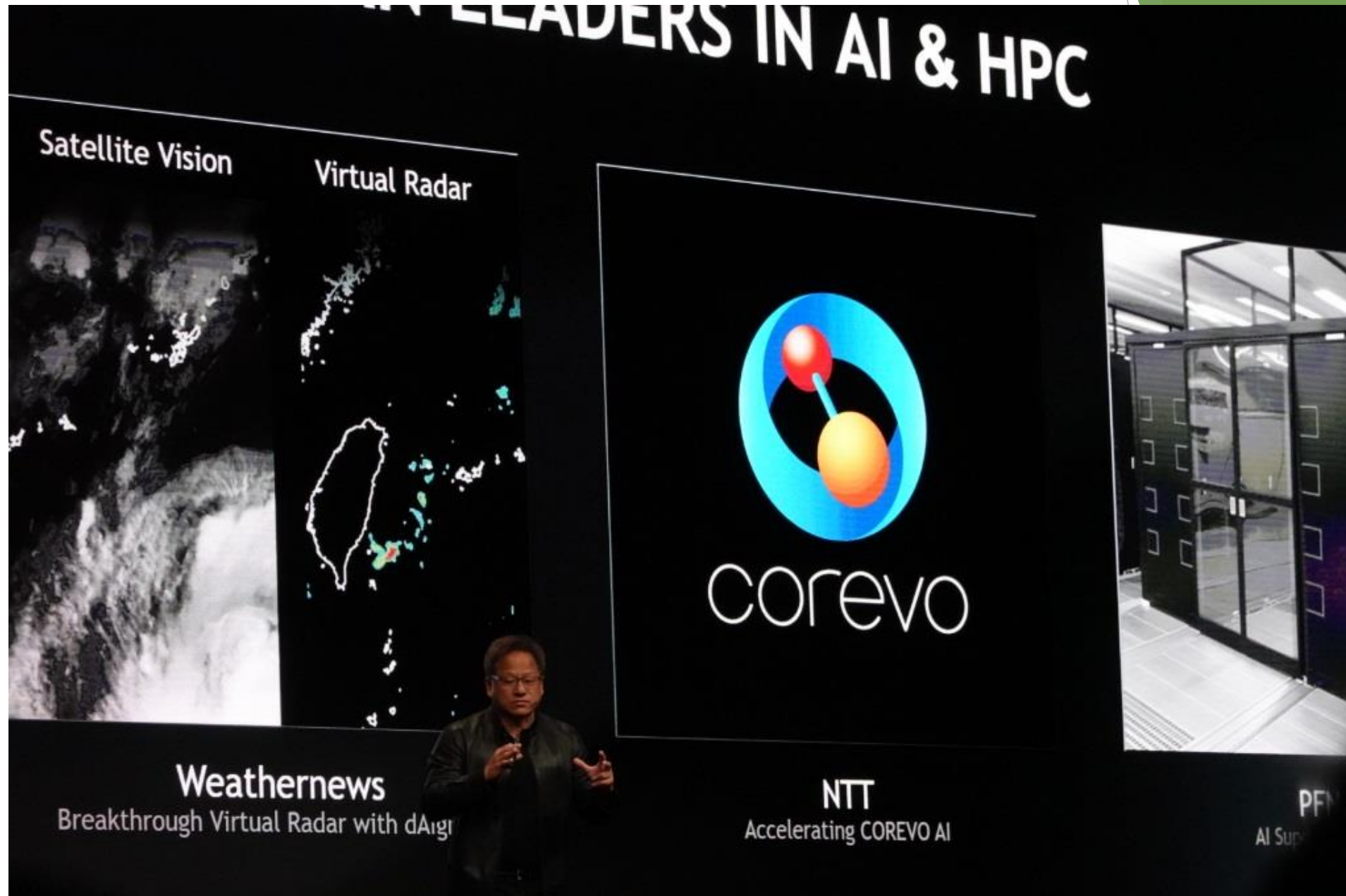
Virtual Radar

Weathernews  
Breakthrough Virtual Radar with dAig

corevo

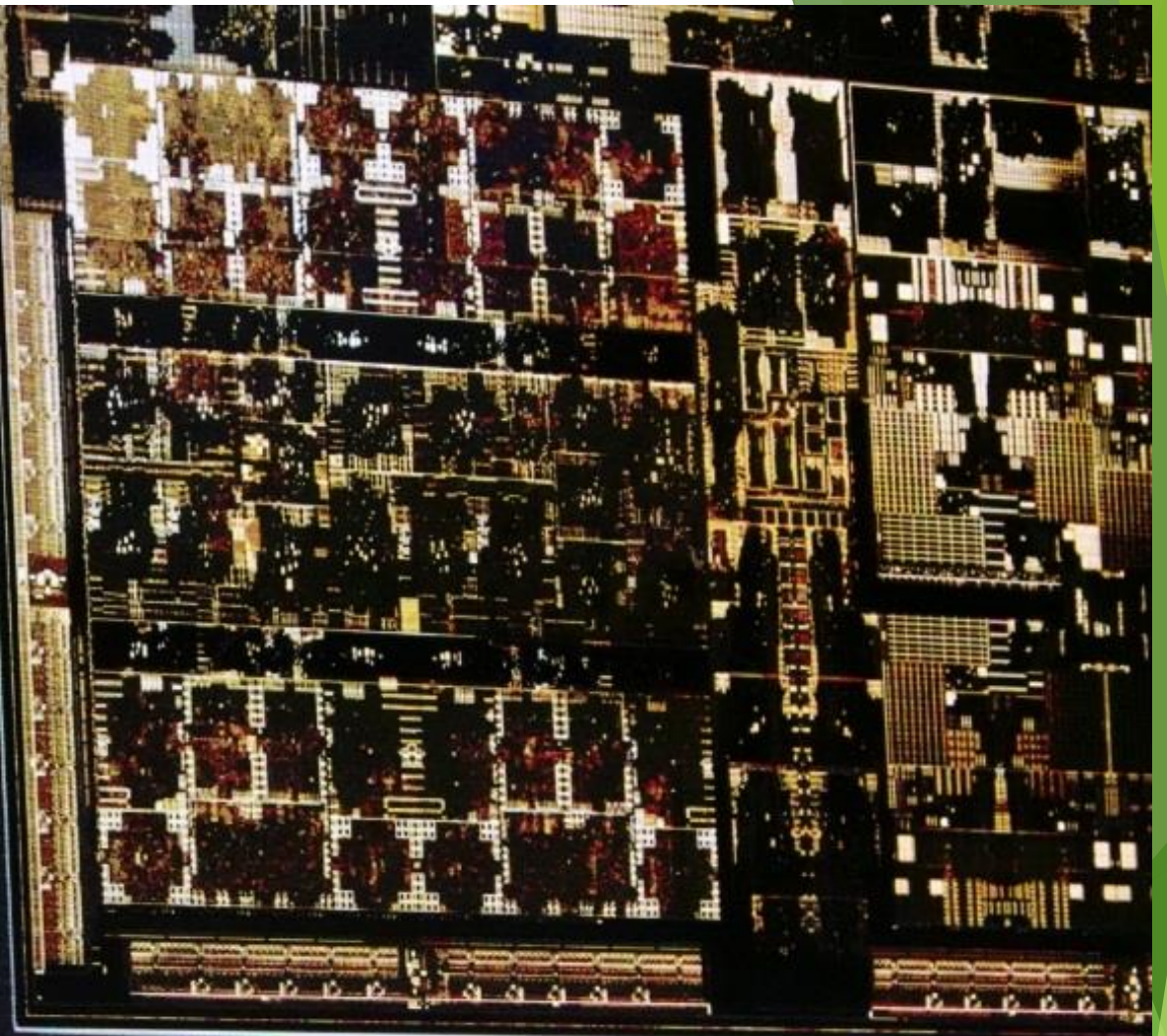
NTT  
Accelerating COREVO AI

PF  
AI Sup



Stereo & Optical Flow Engine  
2x 3.1 TOPS

Volta Tensor Core GPU  
FP32 / FP16 / INT8 Multi-Precision  
512 CUDA Tensor Cores  
2.8 CUDA TFLOPS (FP16)  
22.6 Tensor Core DL TOPS



Nvidia Xavier(エッジコンピューティング) も販売しているよ15万

256-Bit LPDDR4X  
137 GB/s



# Variational Autoencoders for NLP: Particular Difficulties, Recent Solutions, and Practical Applications

Research Scientist, Cogent Labs

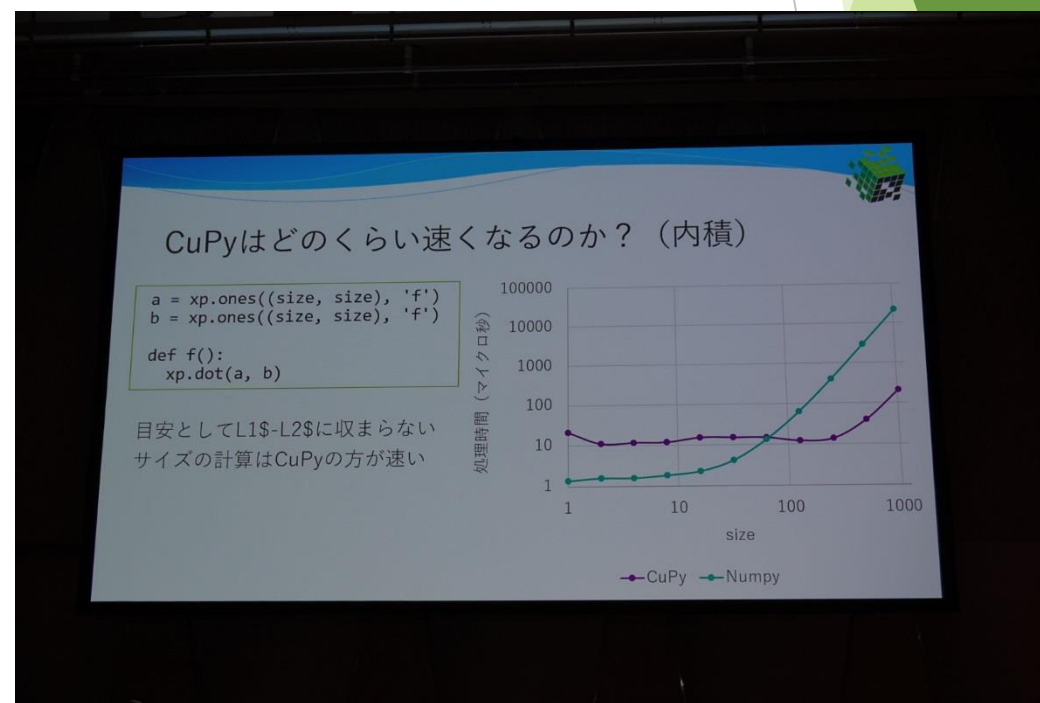
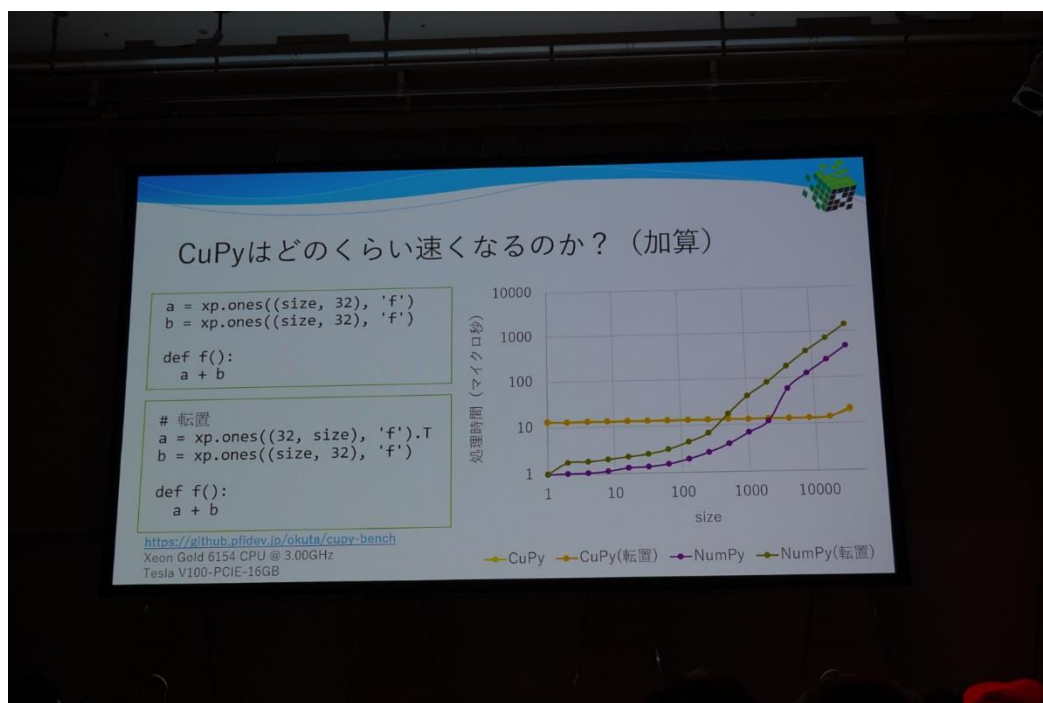
- ▶ 代官山にある超絶おしゃれオフィスであって、AIベンチャーでは有名で、いろんなバックグラウンドのリサーチャー、エンジニアが在籍している COGENT LABS のリサーチャーの発表。
- ▶ Variational Autoencoder(VAE) をテキストマイニングへ応用した例および、VAEの概論について説明されていました。VAEはこの辺をよめばいいです。（ <https://qiita.com/kenchin110100/items/7ceb5b8e8b21c551d69a> ）
- ▶ DiSAN についても言及されてました。論文はこちら  
<https://arxiv.org/abs/1709.04696>
- ▶ これらの技術を利用したプロダクトとして Kaidoku という文書検索システム。クラスタリングや時系列表示ができるというも以下の会社や団体ですでに利用しているらしい。
  - ▶ 野村證券  
リサーチャーが情報の整理などレポートの作成を支援する形で利用されているということでした。作業スピードが90[%]削減されたとのこと。
  - ▶ 鎌倉市  
市政などの評判分析的なものをtwitterのデータでやっている。  
twitterの分析はいろんな崩れた表現、短文等で難しい。



# CuPy -NumPy 互換 GPU ライブラリによる Python での高速計算-

- ▶ PFNの中の人講演、chainerで利用されているGPUをnumpyのように利用できるCuPyのあれこれの発表
- ▶ 基本的にnumpyと同じIFと動作になるように心がけているが、GPU利用時のボトルネック（GPUメモリ転送、起動）などがあるので小さい規模のデータ処理では速度があたりまえだが、numpyにまけているのだが、それを速くする努力はしている。
- ▶ Numpyで実装されているがCupyで実装されていない関数などの追加などががんばっている。
- ▶ GPUメモリがたりない場合は（UMAを使えるよ、性能はアレだどうしても動かしたいときによい）

<https://docs-cupy.chainer.org/en/stable/reference/generated/cupy.cuda.MemoryPool.html>



気になたポスター（ほとんど見れてない）

## Performance evaluation of multi-GPU implementation of RI-MP2 method on Tesla V100 GPUs

**Michio Katouda\***

**Acknowledgement:** Grants-in-Aid for Scientific Research from JSPS (ID: 15K17816).

Amazon Elastic Compute Cloud (EC2),

Research Institute for Information Technology, Kyushu Univ.

## Abstract

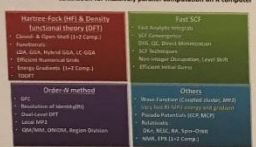
We have ported the multi-GPU massively parallel implementation of resolution-of-identity second-order Møller-Plesset perturbation (RI-MP2) energy calculation into for supercomputers where Tesla V100 GPUs are installed such as AI Bridging Cloud Infrastructure (ABCI) at Advanced Industrial Science and Technology (AIST) and P3 instance of Amazon Elastic Compute Cloud (EC2). In this presentation, we report the overview of implementation and the results of performance evaluation of the implementation. We also discuss the performance comparison between the results using Tesla V100 GPUs on p3.xlarge instance of Amazon EC2 and those using Tesla K40 and P100.

## Summary

- Multi-GPU implementation of RI-MP2 method has been ported to Tesla V100 system such as ABCI@AIST and Amazon EC2.
- Speed ups of calculation are attained using V100 GPUs of p3.xlarge instance@Amazon EC2 compared with the CPU only calculations.
- Performances using p3.xlarge are slightly worse than those using P100 system in ITO@Kyushu Univ. but acceptable for product jobs.

Introduction: NTChem/RI-MP2 program for electronic structure calculation of large molecular systems

**NTChem** A comprehensive new software of *ab initio* electronic structure calculation for massively parallel computation on K computer.



Theories and methods for electronic structure calculations of large molecules on massively parallel supercomputers are available  
Webpage: [http://molsc.riken.jp/ntchem\\_e.html](http://molsc.riken.jp/ntchem_e.html)

## Resolution-of-identity second order Møller-Plesset perturbation (RI-MP2) method

### MP2 theory

- Treating electron correlations based on Møller-Plesset perturbation theory with truncation of perturbation expansion term in second order  $E^{(2)}$
- The simplest method for calculations of electron correlation in ab initio molecular orbital theory
- Robust method for calculations of non-covalent inter-molecular interactions
- High computational costs:  $O(N^4)$  unlike Hartree-Fock theory and density function theory:  $O(N^3)$
- High storage requirements:  $O(N^2) \rightarrow O(N^3)$

Total MP2 energy	MP2 correlation energy	6-center (4c) NVD 2-electron repulsion integrals
$E = E^{(0)} + E^{(2)}$	$E^{(2)} = \frac{1}{2} \sum_{ab} \sum_{cd} \frac{\langle ab g cd \rangle [2 \langle ab g ab \rangle - (\langle ab g ba \rangle)]}{\epsilon_a + \epsilon_b - \epsilon_c - \epsilon_d}$	$\langle ab g cd \rangle = \sum_{\mu} C_{\mu a} \sum_{\nu} C_{\nu b} \sum_{\rho} C_{\rho c} \sum_{\sigma} C_{\sigma d} \langle \mu\nu \rho\sigma \rangle$

RI-MP2 method

- This approximation considerably reduces computational costs and resource requirements and applicable to large molecules and suitable for calculations of large molecules
- $$\langle \psi | \hat{H} | \psi \rangle = \sum_i \langle \psi | \hat{H} | \psi_i \rangle \langle \psi_i | \psi \rangle = \sum_i \langle \psi | \hat{H} | \psi_i \rangle \sum_j \langle \psi_j | \psi \rangle \langle \psi_j | \psi_i \rangle$$

### Computational scheme of RI-MP2 energy calculation

Isotop: molekularer gewicht: atomgewicht (A/G) molgewicht (M/G) 100g

Molecular orbital (MO) coefficients $c_{\mu}^i$ and orbital energies $\epsilon_i$ , obtained by HF calculations	
Step 1 3-center MO 2-electron integral (1)	$(iij) = \sum_{\mu, \nu, \lambda} c_{\mu}^i c_{\nu}^j c_{\lambda}^i \langle \mu \nu \lambda   iij \rangle$ $[iij] = \sum_{\mu, \nu, \lambda} c_{\mu}^i c_{\nu}^j c_{\lambda}^i \langle \mu \nu \lambda   iij \rangle$
Step 2 3-center MO 2-electron integral (2)	$K^i = \sum_{\mu, \nu} (i\mu\nu)   \mu \nu \rangle \langle i  $
Step 3-1 4-center MO 2-electron integral	$(i\mu\nu\lambda) = \sum_{\mu, \nu, \lambda} c_{\mu}^i c_{\nu}^j c_{\lambda}^i \langle \mu \nu \lambda   i\mu\nu\lambda \rangle$
Step 3-2 Output: MP2 correlation energy $E^{(2)} = \sum_{i,j} \frac{(i\mu\nu\lambda) [2(i\mu\nu\lambda) - (i\mu\nu\lambda)]}{\epsilon_i + \epsilon_j - \epsilon_{\mu} - \epsilon_{\nu} - \epsilon_{\lambda}}$	$E^{(2)} = \sum_{i,j} \frac{(i\mu\nu\lambda) [2(i\mu\nu\lambda) - (i\mu\nu\lambda)]}{\epsilon_i + \epsilon_j - \epsilon_{\mu} - \epsilon_{\nu} - \epsilon_{\lambda}}$

Operation: O(N)<sup>3</sup>  
Memory: O(N)<sup>2</sup>  
Difficult to scale in 7 mode

Computational bottleneck step  
Operation: O(N)<sup>3</sup>  
Memory: O(N)<sup>2</sup>  
Difficult to scale in 1 mode

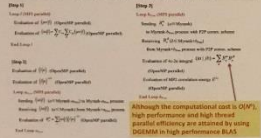
Operation: O(N)<sup>3</sup>  
Memory: O(N)<sup>2</sup>

Computation costs are relatively expensive than HF calculation  
Development of efficient algorithm and implementation is desired for application to real molecular systems

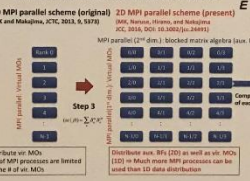
## Two dimensional (2D) hierarchical MPI parallelization and GPGPU implementation of RI-MP2 energy calculation

## Original MPI/OpenMP hybrid parallel algorithm

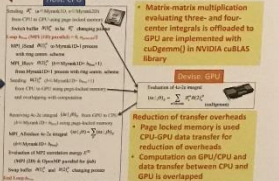
- MPI parallelization: outermost loops. (Miki and Miki@na, JCTC, 2012, 9, 5373)
- The virtual MOs that are much larger than occupied MOs are used for MPI parallelization. This considerably enhances the parallel efficiency.
- Tasks inside of MPI parallelized loops are parallelized by OpenMP.



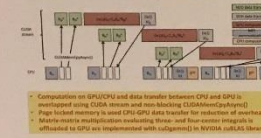
### 2D hierarchical MPI parallelization scheme



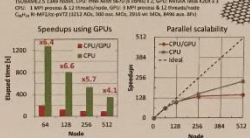
Host: CPU GPGPU implementation



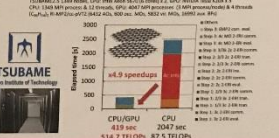
### Reduction of CPU-GPU data transfer overheads



**Speedups and parallel scalability on TSUBAME2.5**

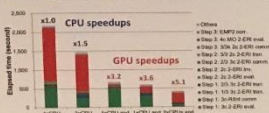


**Largest GPU calculations on TSUBAME2.1**



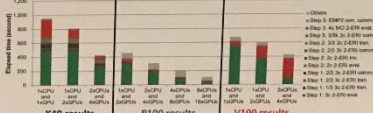
### Performance evaluation of GPGPU implementation using Tesla V100, P100, and K40 on multiple CPU/GPU systems

Performance of Tesla V100 system  
(ge instance of AWS EC2 with NVIDIA GPU cloud)



Good scalability and considerable speedups are attained with multi-GPU computation due to the speed up of DGEMM calculations offloading to multi GPUs.

Comparison of performance of Tesla V100, P100, and K40 systems  
(V100: p3.8xlarge@AWS EC2, P100: ITO@Kyushu Univ., K40@own system)



Using 2 GPUs (and 2 CPUs), Tesla V100@AWS is about 2 times slower than Tesla P100@ITD. Main reason of performance degradations are come from the slow calculations of CPU part. Further analysis of performance data is needed to perform effective tuning.



気になたポスター（ほとんど見れてない）

俳句よむやつあるんやー

