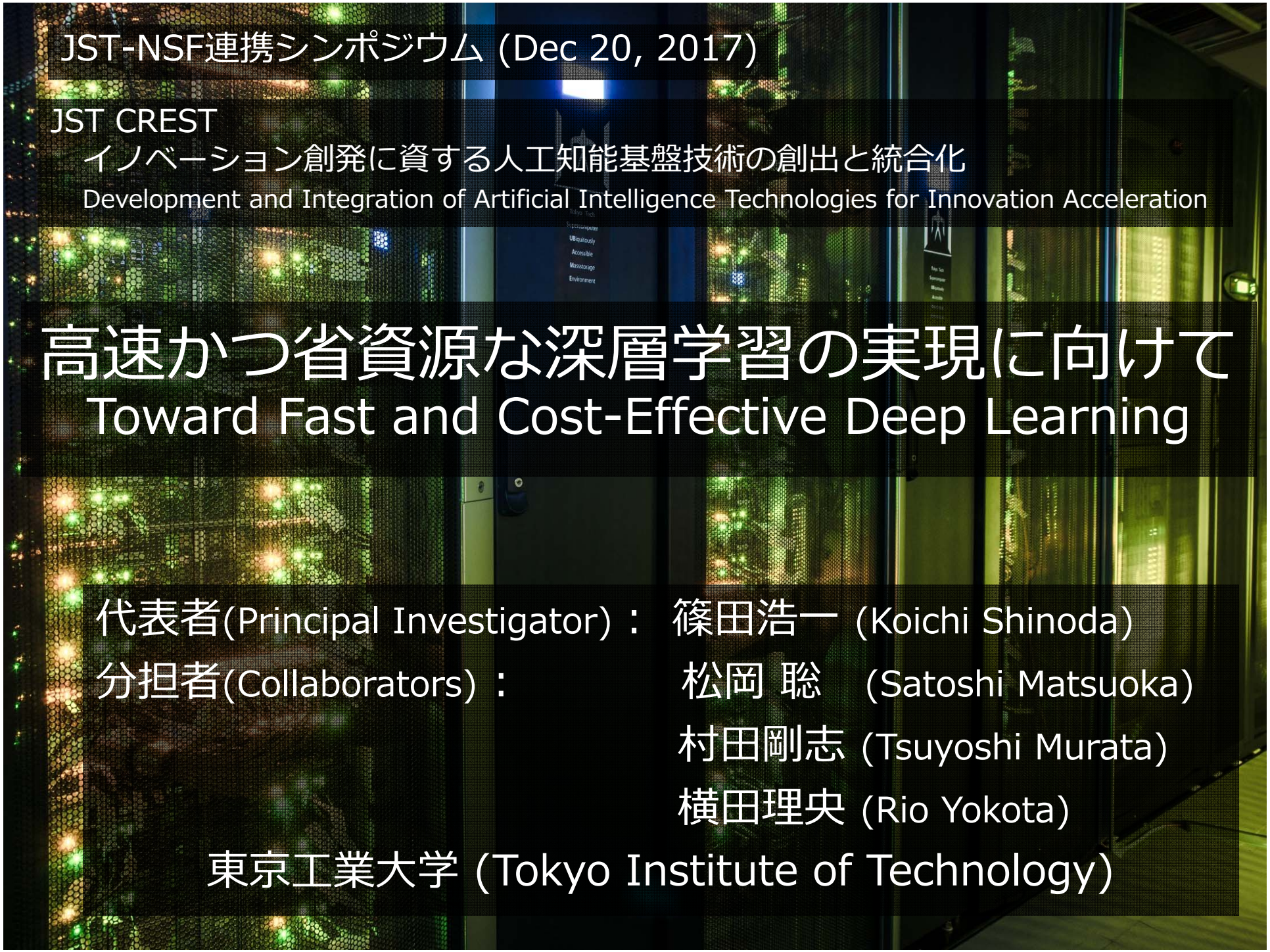


JST-NSF連携シンポジウム (Dec 20, 2017)

JST CREST
イノベーション創発に資する人工知能基盤技術の創出と統合化
Development and Integration of Artificial Intelligence Technologies for Innovation Acceleration

高速かつ省資源な深層学習の実現に向けて
Toward Fast and Cost-Effective Deep Learning

代表者(Principal Investigator)： 篠田浩一 (Koichi Shinoda)
分担者(Collaborators)： 松岡 聡 (Satoshi Matsuoka)
村田剛志 (Tsuyoshi Murata)
横田理央 (Rio Yokota)
東京工業大学 (Tokyo Institute of Technology)

目的 (Motivation)

- 安全・安心なスマート社会の実現
  Toward safe and secure smart society
- 高度な映像技術の開発
  Development of advanced video recognition technologies
  - 事故を未然に防止 Prevent traffic accidents from happening
  - 異常を早期発見 Detect abnormalities in their early stages
- 人の動きは複雑、予測が難しい
  Human behavior is complex, difficult to be predicted
  → 人工知能で実現できないか？
  Challenge problem for Artificial Intelligence

課題(Problem)

- これまでのターゲット(Target until now)：
  静止画(Static images)、大きい物体(Large objects)



- この研究の対象(Our Target)：
  動画(Video)、小さい物体とその動き(Small objects and their movement)


我々のプロジェクト (Our Project)

- 2016年12月より開始 (1年経過)
  Start from Dec. 2016 (1 year has passed)
- ステージゲート方式 (Stage gate process)
  Small phase (2.3 year) + Large phase (3 year)
- 東工大の4名の研究者 (Four members in TokyoTech)
  - 横田 (Yokota)
  - 松岡 (Matsuoka)

  } 高性能計算(HPC)

  - 篠田 (Shinoda)
  - 村田 (Murata)

  } 機械学習(Machine learning)
- 産業総合技術研究所と共同研究
  (Collaborate with Advance
  Institute of Science and Technology)

AIST-Tokyo Tech
Real World Big-Data Computation
Open Innovation Laboratory
(RWBC-OIL)

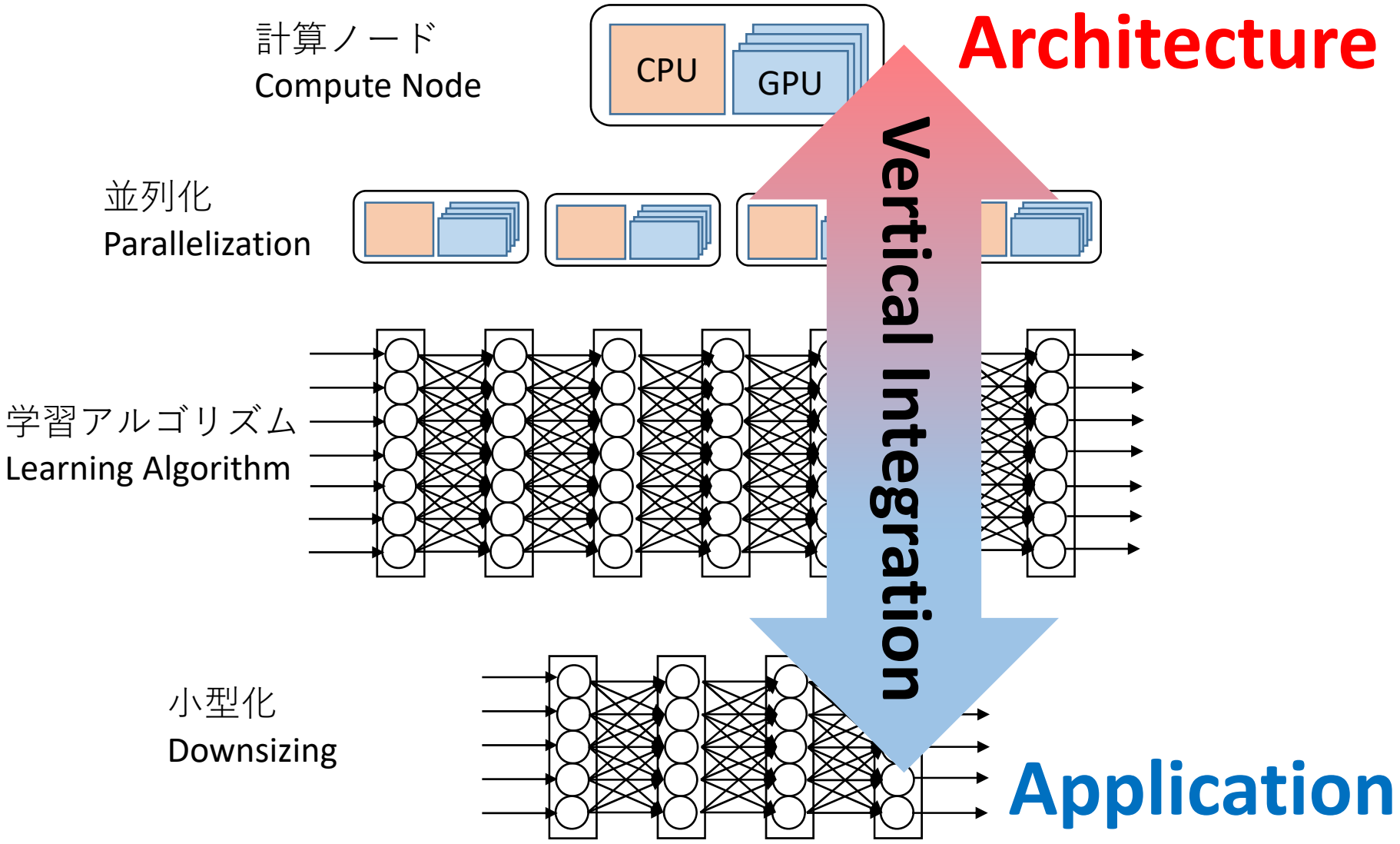
課題 (Problem)

1. 大量の画像の実時間での解析
   Analyze a huge amount of images in real-time
2. 環境の変化に速やかに適応
   Rapidly Adapt to the changes in environmental conditions
3. 端末側での計算 → 通信量の削減
   Edge Computing　　Reduce traffics on Internet

これらの課題は密接に関連
These problems are deeply related with each other

→ 同時に最適化
   Simultaneous optimization

Approach – Co-Design –



1000x speed by 1/1000 memory

研究計画 (Research Plan)

Small Phase: Co-Designにおける要素技術を開発
(Develop each component in the Co-Design framework)

- 計算ノード(Computer Node)
- 並列化 (Parallelization)
- 学習アルゴリズム (Learning algorithm)
- 小型化 (Downsizing)

Large Phase: 統合評価 (Integration & Evaluation)

- 実時間動作 (Real time operation)
- 実環境で評価 (Evaluate in real applications)
  - e.g. Smart City Sensor
    - Argonne National Lab & Chicago Univ.
- オープンプラットフォーム化 (Open platform)
  - APIやツールキットを提供
    (Release API, Tool kit on Cloud)

スマートシティセンサー
Smart City Sensor

Goal in Small Phase

Component	Speed	Memory
Compute node	50x	1/10
Parallelization	10x	
Learning Algorithm	10x	1/10
Downsizing		1/100
Total	> 1000x	< 1/1000

NIST TRECVID
マルチメディア検出(Multimedia Event Detection)

- ビデオから複雑な「イベント」を検出
  Detect a complex “event” from a video clip

  e.g. “ホームラン(Batting a run in)”, “ケーキ作りMaking a cake”
- インターネットビデオ(Consumer video) 7,879 hour

2017年の成果

What we have done in year 2017

TokyoTech Supercomputer TSUBAME 3.0 released

計算ノード (Compute Node)

自然勾配法は最急降下法に比べ1/100のステップ数で収束するが1ステップに100倍の計算時間がかかるので計算時間は同じ

Natural Gradient converges 100x faster
but takes 100x per iteration

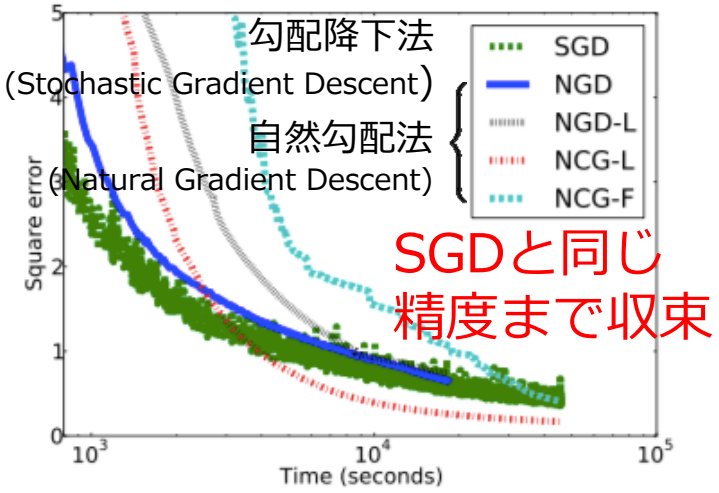
If we can accelerate the Fisher Matrix Calculation
we can reduce the time per iteration to 1/100

自然勾配法のボトルネックであるFisher情報行列の計算を高速化することで1ステップあたりの計算時間を短縮し全体の計算時間は最急降下法の1/100に

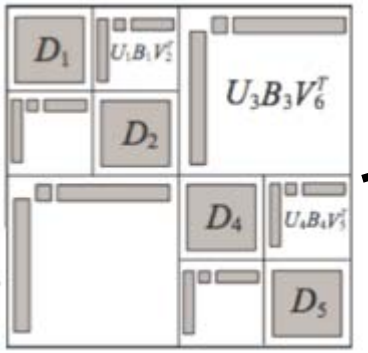
$$L(\theta + \Delta\theta) = L(\theta) + G^{-1}\frac{\partial L}{\partial \theta}\Delta\theta$$

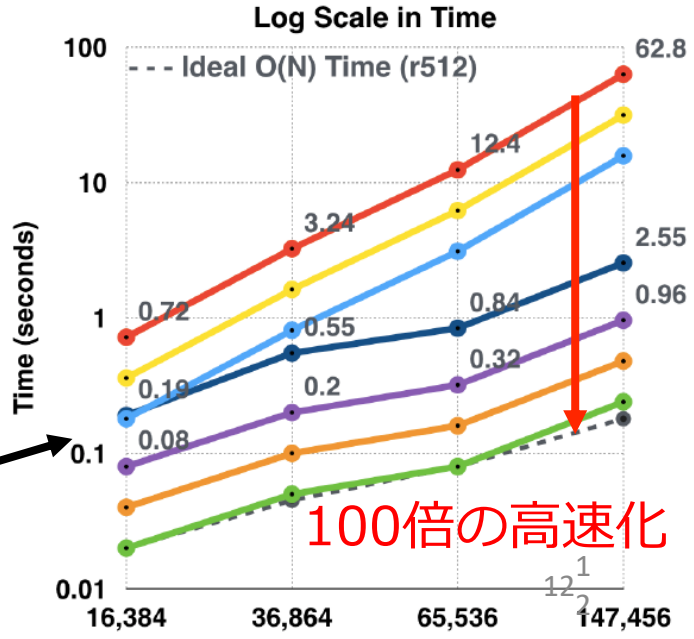
Fisher Matrix

Hierarchical
Low-rank
approximation



SGDと同じ
精度まで収束

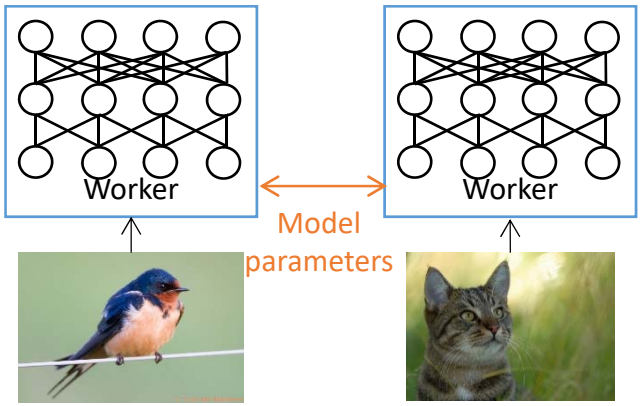




100倍の高速化

並列化(Parallelization)

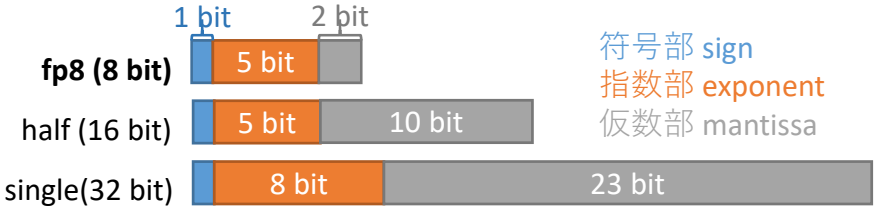
データ並列分散学習 (Data-parallel Distributed Deep Learning)



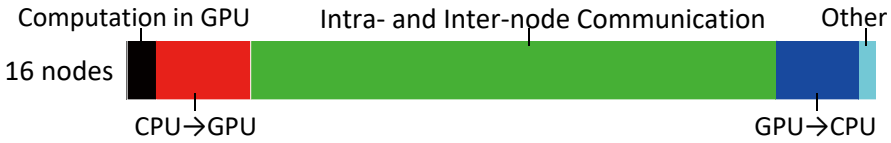
同じネットワークを**別々のデータで**学習
Train the same network with **different data**
パラメータ同期のための通信が必要
Parameter must be synchronized between workers

縮退精度通信 (Communication in Reduced Precision)

通信時間が並列化時のボトルネックに：8ノード使用時も1ノードの**3.59倍**しか速くならない
Communication time is bottleneck in parallelization: Only **3.59x** achievable in 8 nodes



実行時間のうちGPU演算は**3.9%** Computation is only **3.9%**
(CaffeNet on TSUBAME-KFC/DL)



GoogLeNetの学習で8ノードを用いた場合に1ノードの7.67倍の高速化を達成
7.67x speedup on 8 nodes compared to 1 node, in GoogLeNet Training

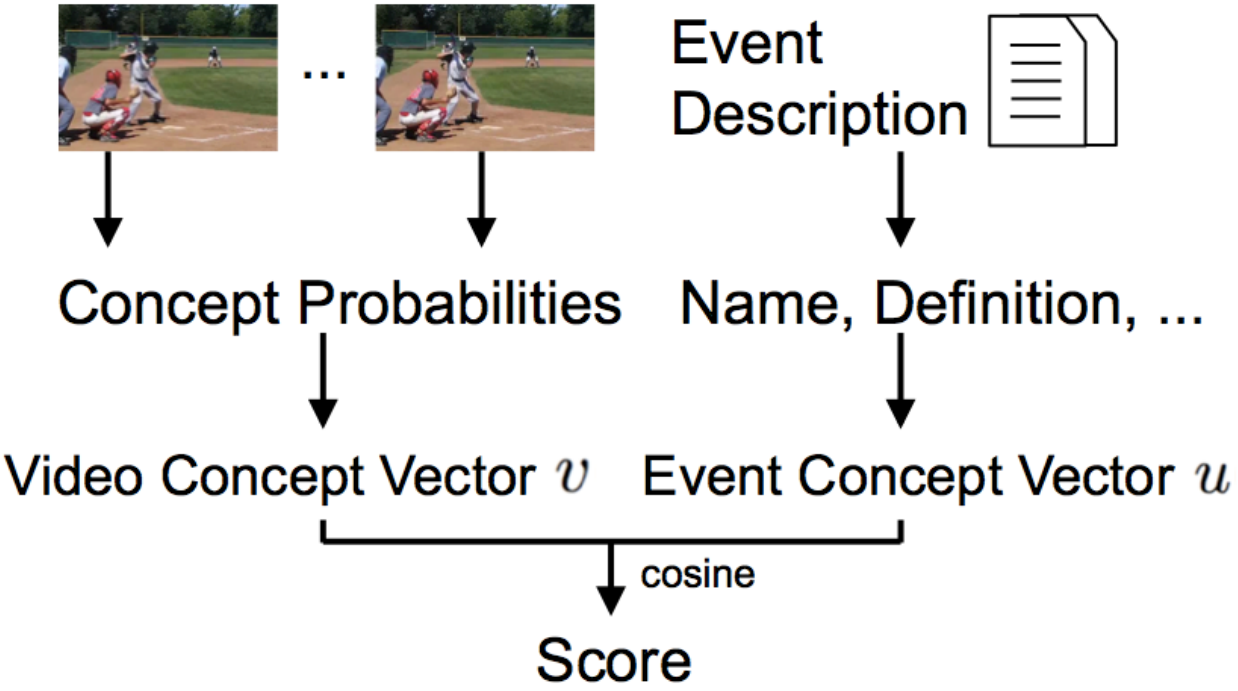
学習アルゴリズム(Learning algorithm)

データへのラベル付与工数の削減(Reduce the cost of data annotation)

→半教師付き学習(semi-supervised learning)

少量の教師ありデータと大量のラベルなしデータを使用
(Use a small amount of annotated data and a large amount of data without annotation)

A hybrid of supervised and zero-shot classifiers



小型化(Downsizing)

- モバイル利用のためのDNNサイズ圧縮(Compressing DNN for mobile devices)

手法	圧縮率	最上位の精度	上位5つの精度	サイズ
Methods	**Comp. Rate**	**Top-1 Acc.**	**Top-5 Acc.**	**Size (MB)**
Original	-	0.58	0.80	240
Knoll, 2012 (P+H)	38x	0.58	0.80	6.3
Han, 2016 (P+Q+H)	35x	0.58	0.80	6.9
Zhou, 2017 (P+Q)	89x	0.58	0.80	2.69
Ours (P+Q+D)	90x	0.68	0.89	2.64

P：枝刈り(Pruning)
Q: 量子化(Quantization)
H: 符号化(Huffman Encoding)
D: 符号化(DEFLATE)

特別なハードウエア不要で1/90の圧縮率を実現
(achieved 90x comp. rate without special hardware)

Our achievement in 2017

Component	Speed	Memory
Compute node	7.4x (50x)	1/15(1/10)
Parallelization	11.6x*(10x)	
Learning Algorithm	11.6x*(10x)	2*(1/10)
Downsizing		1/90(1/100)
Total	> x1000	< 1/1000

* : Achievement obtained by the joint work of the two groups

Next year plan (来年の計画)

- 引き続き4つの要素技術を開発、目標達成を目指す (Continue the development in the four components to achieve the goal)
- 高速かつ省コストな深層学習のプラットフォームを構築(Build a unified platform for fast and efficient deep learning)
- Benchmark: NIST TRECVID Two tasks
  - Deep Intermodal Video Analytics (DIVA)

    https://www.iarpa.gov/index.php/research-programs/diva
  - Active Interpretation of Disparate Alternatives (AIDA)

    http://intelligencecommunitynews.com/darpa-aida-program-aims-to-make-sense-of-big-data/
  - Collaboration with NTU (Singapore)