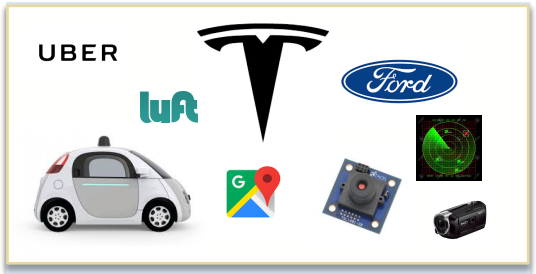
RoCEv2
-RDMA over Converged Ethernet-

Shishio Tsuchiya
shtsuchi@arista.com

ARISTA

幅広いAIの活用領域



自動車

公共＆軍事

クラウドプロバイダー

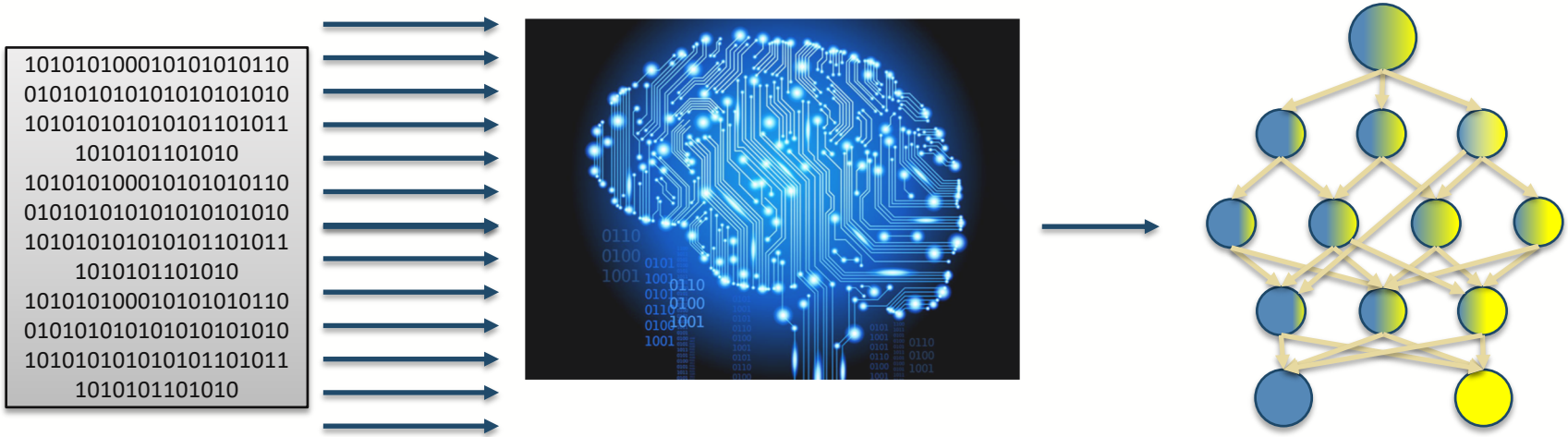
ヘルスケア

金融

消費者

ARISTA

AIとDeep Learning

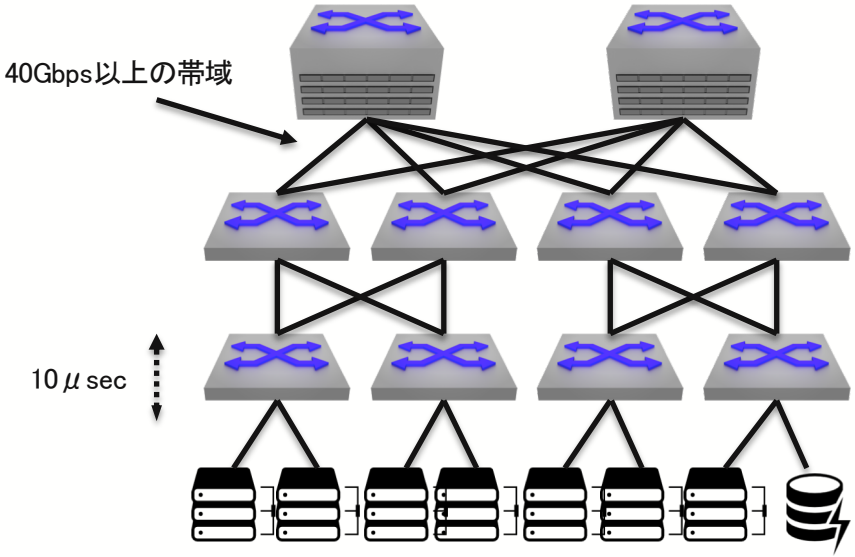


インプットデータ → 深層学習 マルチレイヤートレーニング → 出力/グラフ

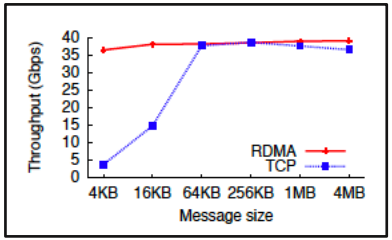
ARISTA

Congestion Control for Large-Scale RDMA Deployments

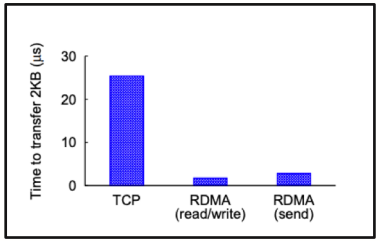
https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf



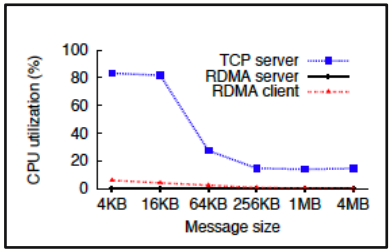
TCP vs RDMA



スループット



転送遅延時間



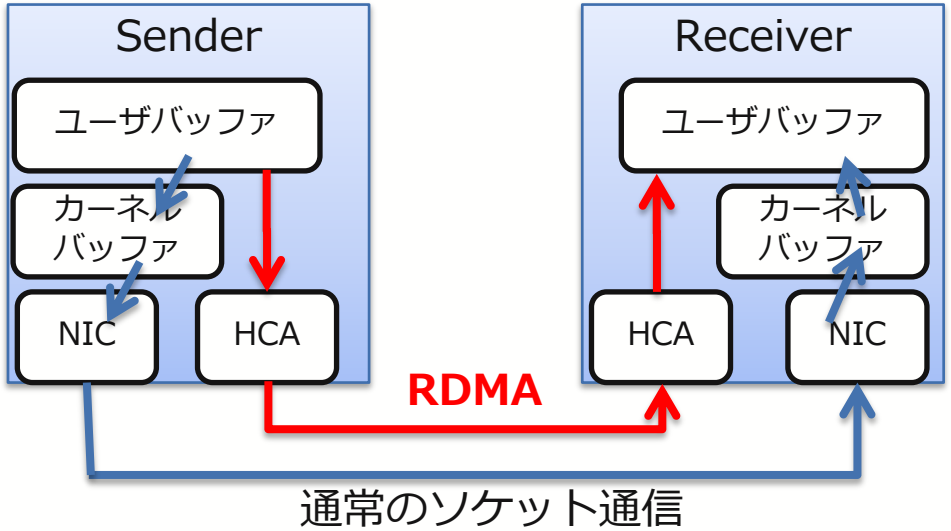
CPU利用率

- 2015年のSIGCOMMの論文
- 昨今のデータセンターにおけるアプリケーション要求は広帯域(40Gbps以上)/超低遅延(10μ秒/hop)/少ないCPU負荷で有ることが求められる

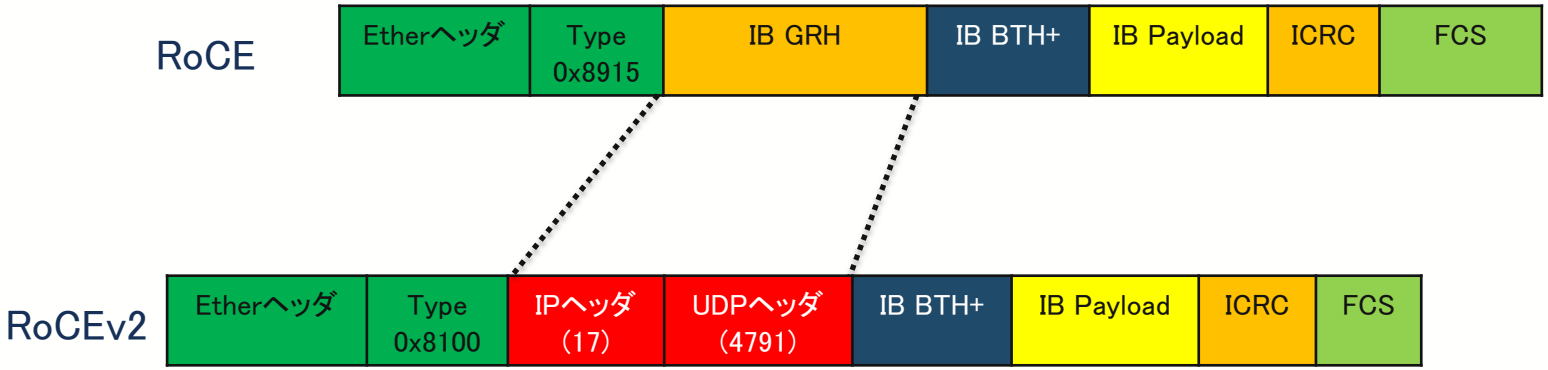
ARISTA

RDMAとは？

- Remote Direct Memory Access(RDMA)
  - アプリケーション間で直接通信
  - システムバス、CPUの負荷を低減
- ロスレスが前提
  - 従来InfiniBandで活用
  - TCPを利用しない=ロスレスが要件
  - RDMA over Ethernet

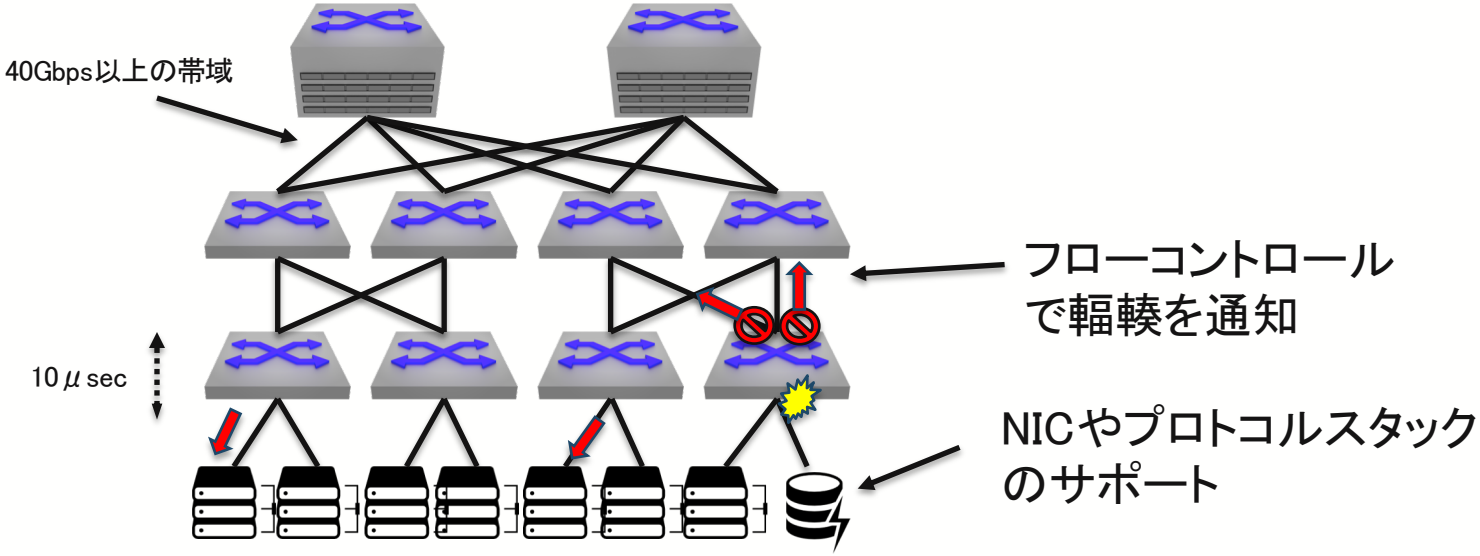


RoCEv2



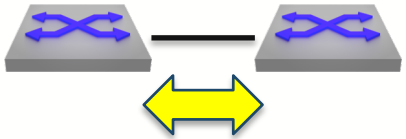
- RoCEはレイヤー2のみの実装であったのに比較して、RoCEv2はIPもしくはIPv6を使用し、ルーティング可能
  UDPポート4791を使用

ロスレスイーサネットネットワークを構築するには



- DCBX(Data Center Bridging Capability Exchange)や PFC(Priority-Based Flow Control)が必要になる

DCBX(Data Center Bridging Capability Exchange)



DCBXアプリケーション設定

```
switch(config)#dcbx application tcp-sctp 860 priority 5
switch(config)#dcbx application tcp-sctp 3260 priority 5
```

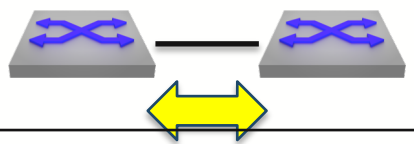
PFC設定

```
switch(config)#interface ethernet 2
switch(config-if-Et2)#priority-flow-control on
switch(config-if-Et2)# priority-flow-control priority 5 no-drop
```

```
switch#show dcbx Ethernet 50
Ethernet50:
  IEEE DCBX is enabled and active
  Last LLDPDU received on Thu Feb 14 12:06:01 2013
  No priority flow control configuration TLV received
  No application priority configuration TLV received
switch#
```

- DCBXはLLDPを通じてアプリケーションやQoSポリシーなどの情報を交換する

DCBX(Data Center Bridging Capability Exchange)



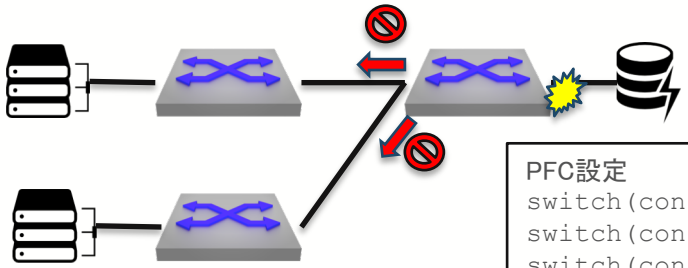
DCBXアプリケーション設定

```
switch(config)#dcbx application tcp-sctp 860 priority 5
switch(config)#dcbx application tcp-sctp 3260 priority 5
```

```
switch#show dcbx Ethernet 50
Ethernet50:
  IEEE DCBX is enabled and active
  Last LLDPDU received on Thu Feb 14 12:08:29 2013
  - PFC configuration: willing
    not capable of bypassing MACsec
    supports PFC on up to 4 traffic classes
    PFC enabled on priorities: 5 7
    WARNING: peer PFC configuration does not match the local PFC configuration
  - Application priority configuration:
    2 application priorities configured:
      tcp-sctp 860 priority 5
      tcp-sctp 3260 priority 5
```

- PFCを設定していない場合

Priority Flow Control (PFC)



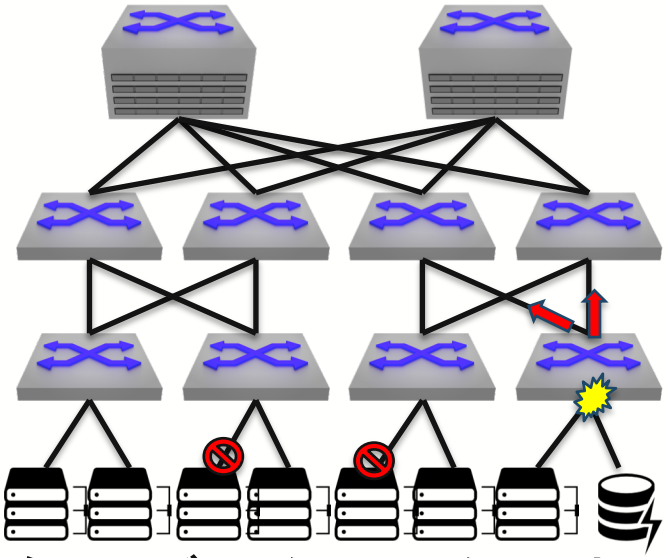
PFC設定

```
switch(config)#interface ethernet 2
switch(config-if-Et2)#priority-flow-control on
switch(config-if-Et2)# priority-flow-control priority 5 no-drop
```

- IEEE 802.1Qbbにて定義、従来のIEEE802.3xフローコントロールの様にPAUSEフレームを送り、トラフィックを制御する
- PFCではトラフィッククラスをVLAN COS値でアプリケーション毎に細かく制御が出来る

- a

ECN – Explicit Congestion Notification



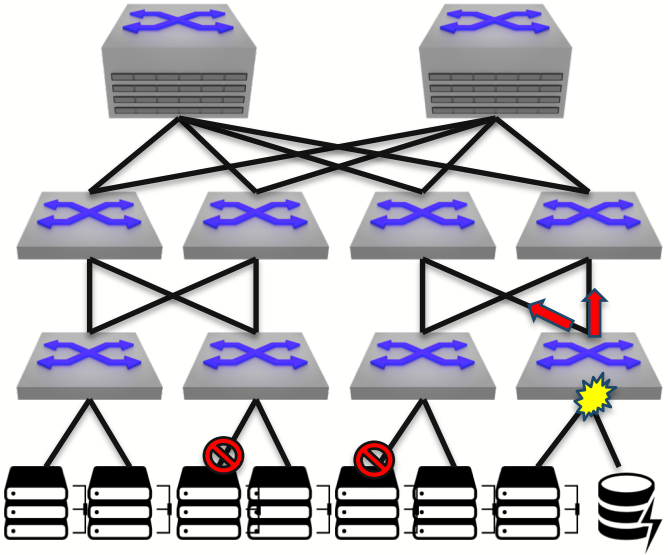
```
   0     1     2     3     4     5     6     7
+-----+-----+-----+-----+-----+-----+-----+-----+
|          DS FIELD, DSCP                 | ECN FIELD |
+-----+-----+-----+-----+-----+-----+-----+-----+
```

Binary	Keyword
00	Not-ECT (Not ECN-Capable Transport)
01	ECT(1) (ECN-Capable Transport(1))
10	ECT(0) (ECN-Capable Transport(0))
11	**CE (Congestion Experienced)**

- 現在のデータセンターネットワークではIP Closが主流
- ECN(RFC3168)を用いてエンドツーエンドに通知する

ECN – Explicit Congestion Notification

```
switch(config)#interface ethernet 3/5/1
switch(config-if-Et3/5/1)#tx-queue 4
switch(config-if-Et3/5/1-txq-4)#random-detect ecn minimum-
threshold 128 kbytes maximum-threshold 1280 kbyte
switch(config-if-Et3/5/1-txq-4)#show active
interface Ethernet3/5/1
   tx-queue 4
      random-detect ecn minimum-threshold 128 kbytes maximum-
threshold 1280 kbytes
switch(config-if-Et3/5/1-txq-4)#
```



- Weighted Random Early Detection (WRED)と共に動作
- maximum thresholdを超えるとECNが開始
  - average queue size = (old_avg * (1-2^(-weight))) + (current_queue_size * 2^(-weight))

まとめ

- 現在のデータセンターアプリケーション要求は下記の通り
  - 広帯域
  - 超低遅延
  - 低CPU負荷
- RCoEv2はIP上で高速/低負荷のRDMAを実現する
- ロスレスアーキテクチャーにはDBEX/PFC/ECNなどが必要