Intel® Gaudi®2 AI Accelerator for Deep Learning Training and Inference

Karthikeyan Vaidyanathan

November 2023

intel®

Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more at https://habana.ai/habana-claims-validation/

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Agenda

- Gaudi2 programming model and recent MLPERF results
- Experience at scale
  - Use case1: SWIFT congestion control
  - Use case2: Packet/message spraying

Intel Xeon and Gaudi2 Processors for Models E2E

Train and deploy large scale GenAI and LLMs

Gaudi2 Clusters and systems for models from billions to trillions of parameters

Fine tune and run thousands of domain specialized models with targeted curated data sets from the data center and the factory floor to devices

Intel Xeon for fine tuning and inferencing models up to tens of billions of parameters

Intel® Gaudi® Accelerator Roadmap

Available Now

GAUDI®

Native RoCE
Scaleup & out

Available via:

- HLS-1 Server (x8)
- SMC Server (x8)
- SDSC
- Public Cloud AWS: EC2

Available Now

GAUDI®2
(7nm)

Native RoCE
Scaleup & out

Available via:

- HLS-Gaudi2 Server (x8)
- SMC Server (x8)
- Aivres/IEI Server (x8)
- Intel Dev Cloud

In Development

GAUDI®3
(5nm)

Native RoCE
Scaleup & out

2024

In Development

Next Generation
AI Accelerator:
Falcon Shores 1

Native RoCE
Scaleup & out

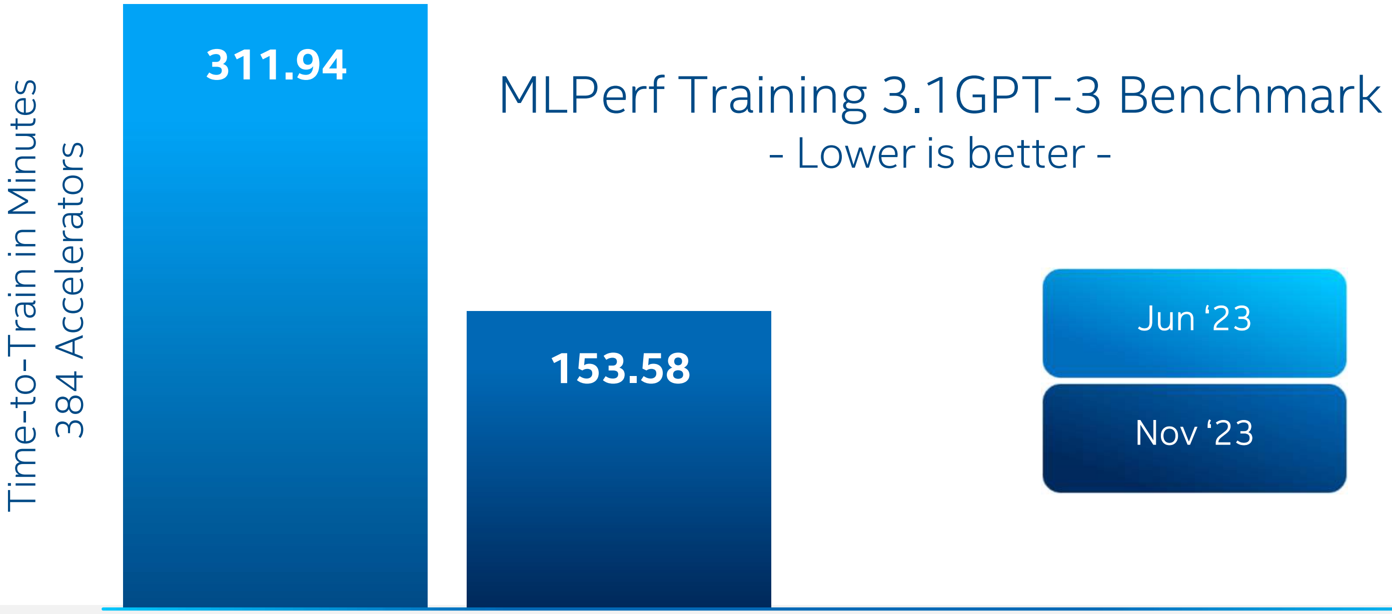
2025

Intel delivers increasingly competitive Training Performance

- One of only three accelerators submitting GPT-3 results: Intel, Nvidia, Google
- Xeon continues to be the only CPU to submit training results on the MLPerf Benchmark.

Intel® Gaudi®2 Accelerator Performance Doubled with FP8

- Intel Gaudi team projected to customers +90% performance gain with FP8
- Delivered more than promised: 103% on GPT-3 industry benchmark

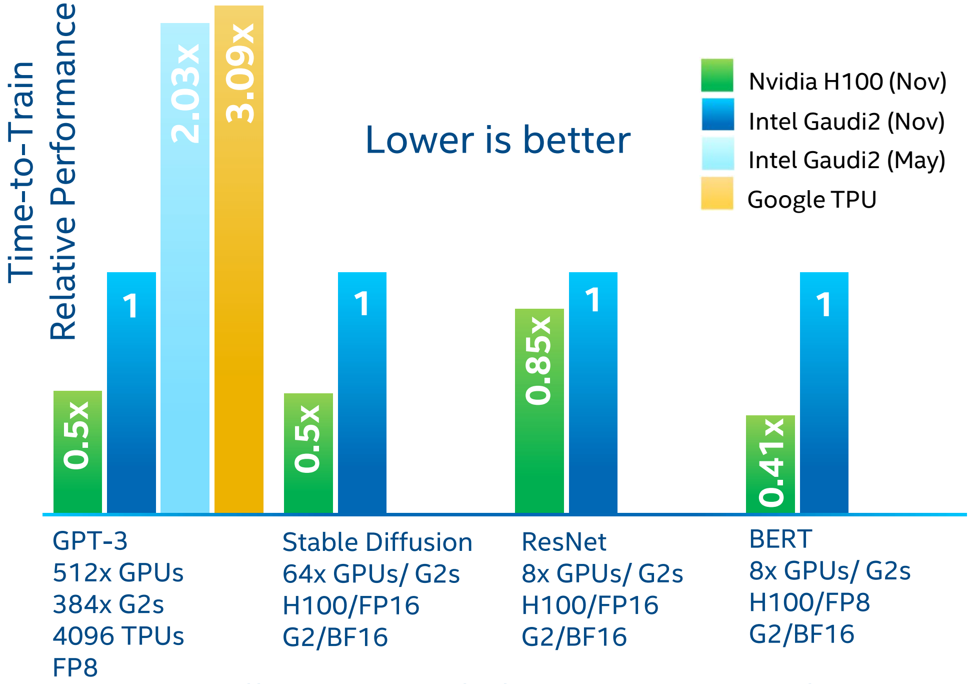


For complete results information and configurations, see MLCommons publication: https://mlcommons.org/en/inference-datacenter-31/

See backup for workloads and configurations. Results may vary.

Intel® Gaudi®2 performance advances strengthen competitive price-performance vs. H100

- Gaudi2 performance on ResNet near that of H100.
- H100 with FP8 outperformed Gaudi2 with BF16 on BERT.
- Vs.TPU, Gaudi2 delivered 3x performance on GPT-3.
- Given its significantly lower server cost vs. H100 server cost, Intel Gaudi2 delivers price-performance advantage vs. H100 across models.



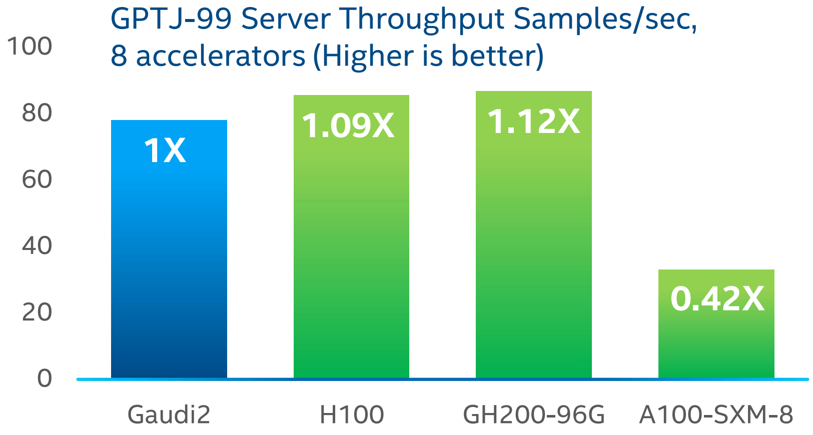
For complete results information and configurations, see MLCommons publication: https://mlcommons.org/en/inference-datacenter-31/

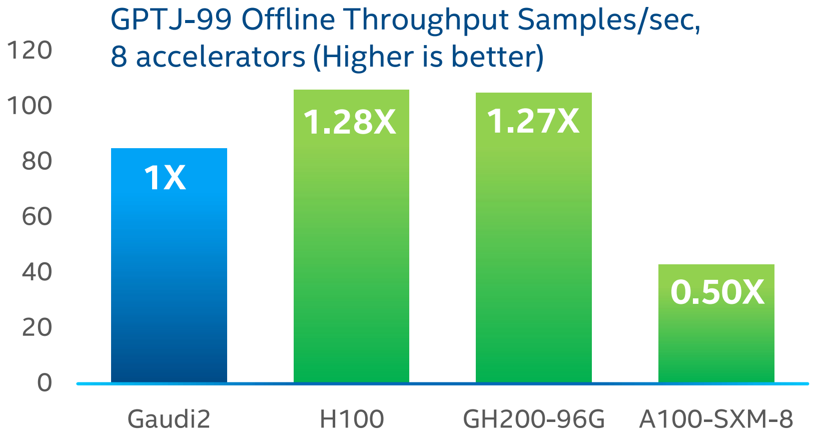
See backup for workloads and configurations. Results may vary.

Outstanding Intel® Gaudi® 2 AI Accelerator performance on MLPerf v3.1 Inference Benchmark

Intel Gaudi2 Accelerator with FP8: near-parity performance on GPT-J (Server) with H100

- Gaudi 2 inference performance on GPT-J: -9% (Server) and -28% (Offline) vs H100
- Gaudi 2 outperformed A100 by 2.4x (Server) and 2x (Offline)
- Gaudi 2 employed FP8 and reached 99.9% accuracy


GPTJ-99 Server Throughput Samples/sec, 8 accelerators (Higher is better)

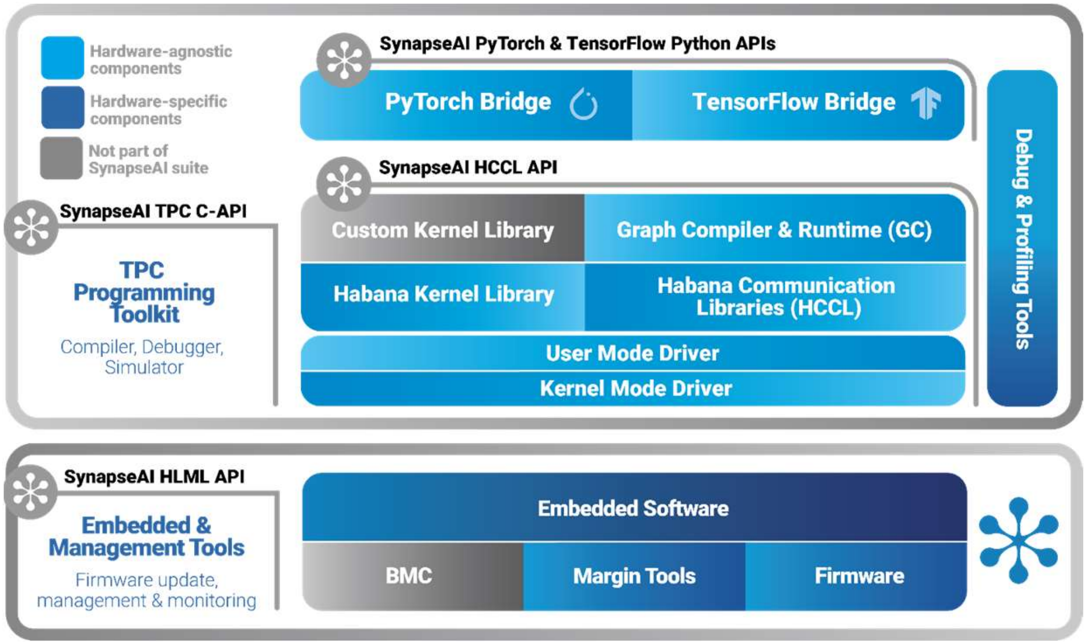

GPTJ-99 Offline Throughput Samples/sec, 8 accelerators (Higher is better)

For complete results information and configurations, see MLCommons publication: https://mlcommons.org/en/inference-datacenter-31/

See backup for workloads and configurations. Results may vary.

SynapseAI Software:
Optimized for Intel® Gaudi® Performance and Ease of Use

SynapseAI®



🟦 Hardware-agnostic components

🟦 Hardware-specific components

⬛ Not part of SynapseAI suite

SynapseAI TPC C-API

TPC Programming Toolkit

Compiler, Debugger, Simulator

SynapseAI PyTorch & TensorFlow Python APIs

PyTorch Bridge | TensorFlow Bridge

SynapseAI HCCL API

Custom Kernel Library | Graph Compiler & Runtime (GC)

Habana Kernel Library | Habana Communication Libraries (HCCL)

User Mode Driver

Kernel Mode Driver

Debug & Profiling Tools

SynapseAI HLML API

Embedded & Management Tools

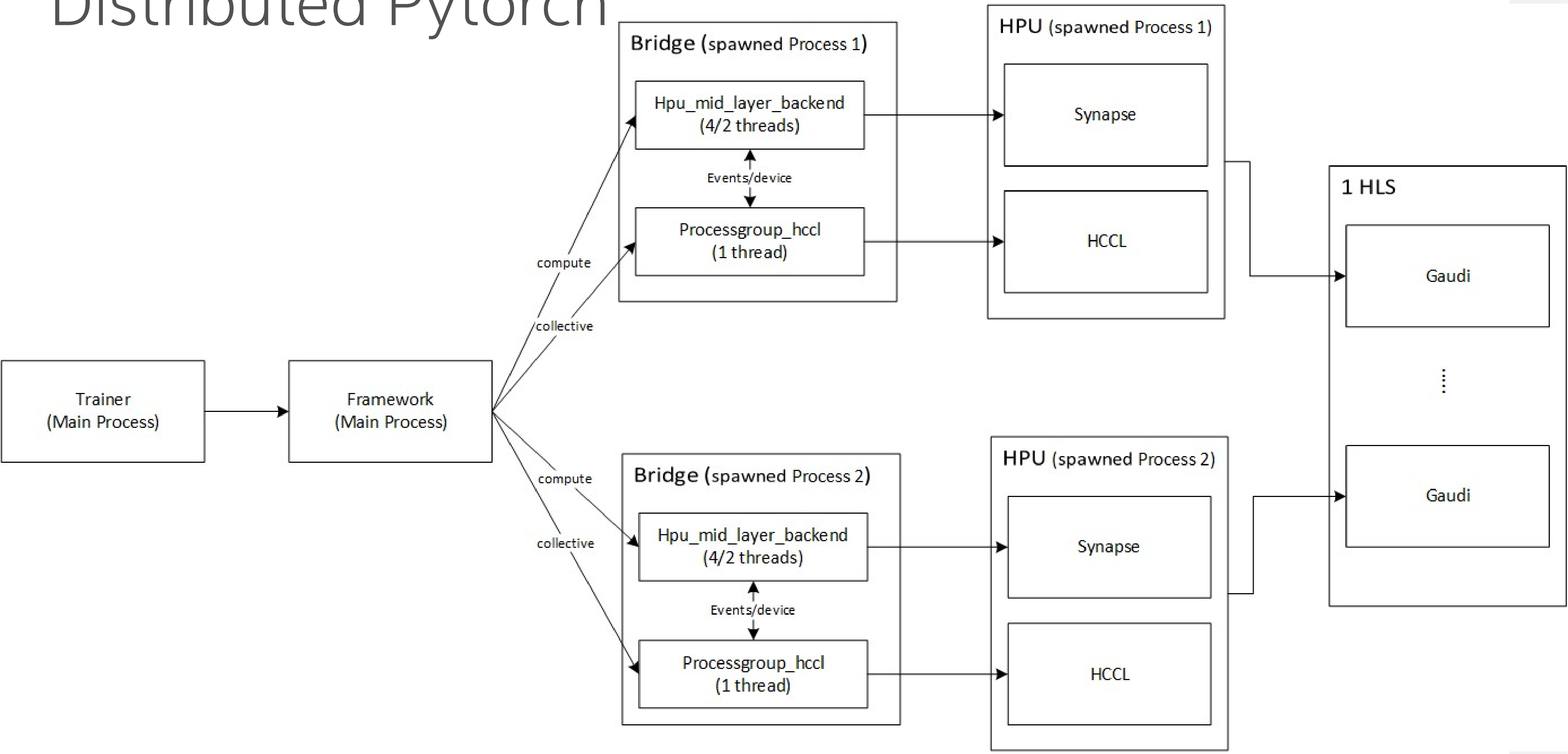
Firmware update, management & monitoring

Embedded Software

BMC | Margin Tools | Firmware

- Shared software suite for training and inference
- Start running on Intel Gaudi accelerators with minimal code changes
- Integrated with PyTorch and TensorFlow
- Rich library of performance-optimized kernels
- Advanced users can write their custom kernels
- Docker container images and Kubernetes orchestration
- Habana Developer Site & HabanaAI GitHub
- Habana Developer Forum

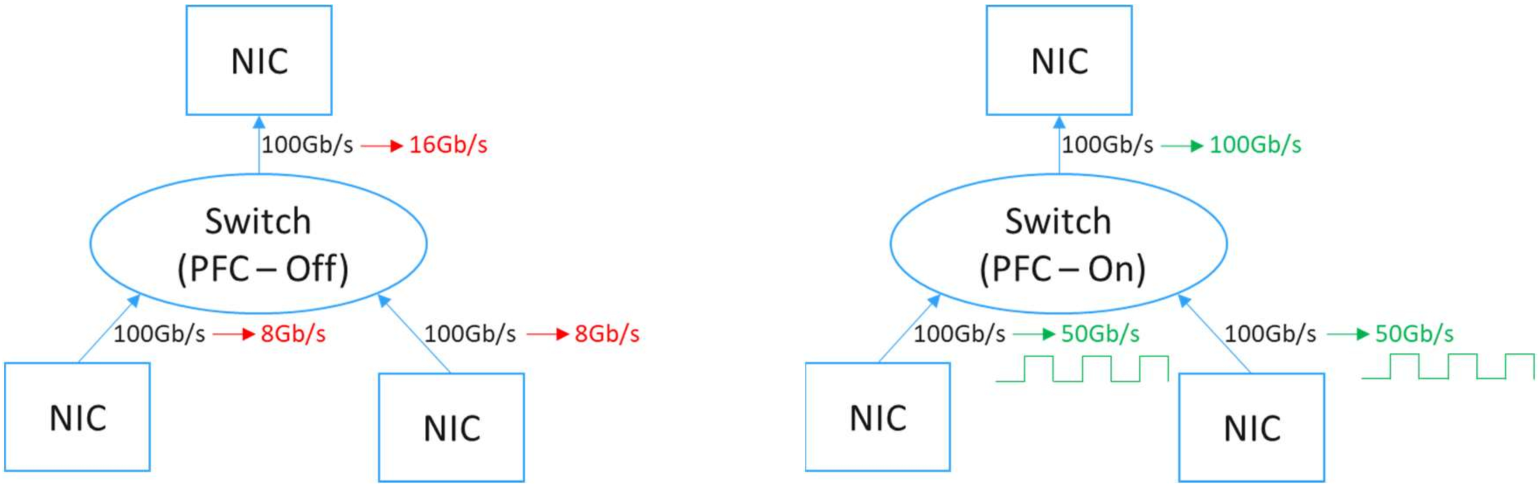
Distributed Pytorch



HCCL API

```
// Communicator creation
hcclGetUniqueId(hcclUniqueId* uniqueId);
hcclCommInitRank(hcclComm_t* comm, int nranks, hcclUniqueId commId, int rank);
// Communicator destruction
hcclCommDestroy(hcclComm_t comm);
// Collectives communication
hcclReduceScatter(void* sbuff, void* rbuff, size_t recvcount, hcclDataType_t datatype, hcclRedOp_t op, hcclComm_t comm, synStreamHandle
stream_handle);
hcclAllReduce(void* sbuff, void* rbuff, size_t count, hcclDataType_t  datatype, hcclRedOp_t op, hcclComm_t comm, synStreamHandle stream_handle);
hcclBroadcast(void* sbuff, void* rbuff, size_t count, hcclDataType_t  datatype, int root, hcclComm_t comm, synStreamHandle stream_handle);
hcclAllGather(void* sbuff, void* rbuff, size_t sendcount, hcclDataType_t datatype, hcclComm_t comm, synStreamHandle stream_handle);
hcclReduce(void* sbuff, void* rbuff, size_t count, hcclDataType_t  datatype, hcclRedOp_t op, int root, hcclComm_t comm, synStreamHandle stream_handle);
hcclAlltoAll(…);
// Point-to-point communication
hcclSend(void* sbuff, size_t  count, hcclDataType_t  datatype, int peer, hcclComm_t comm, synStreamHandle stream);
hcclRecv(void* rbuff, size_t count, hcclDataType_t datatype, int peer, hcclComm_t comm, synStreamHandle stream);
// Aggregation/Composition
hcclGroupStart();
hcclGroupEnd();
```

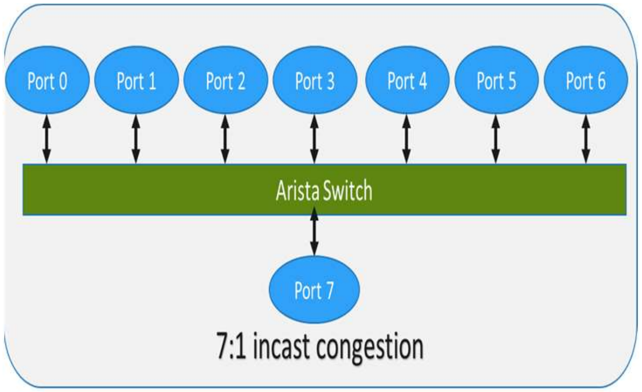
Incast Congestion

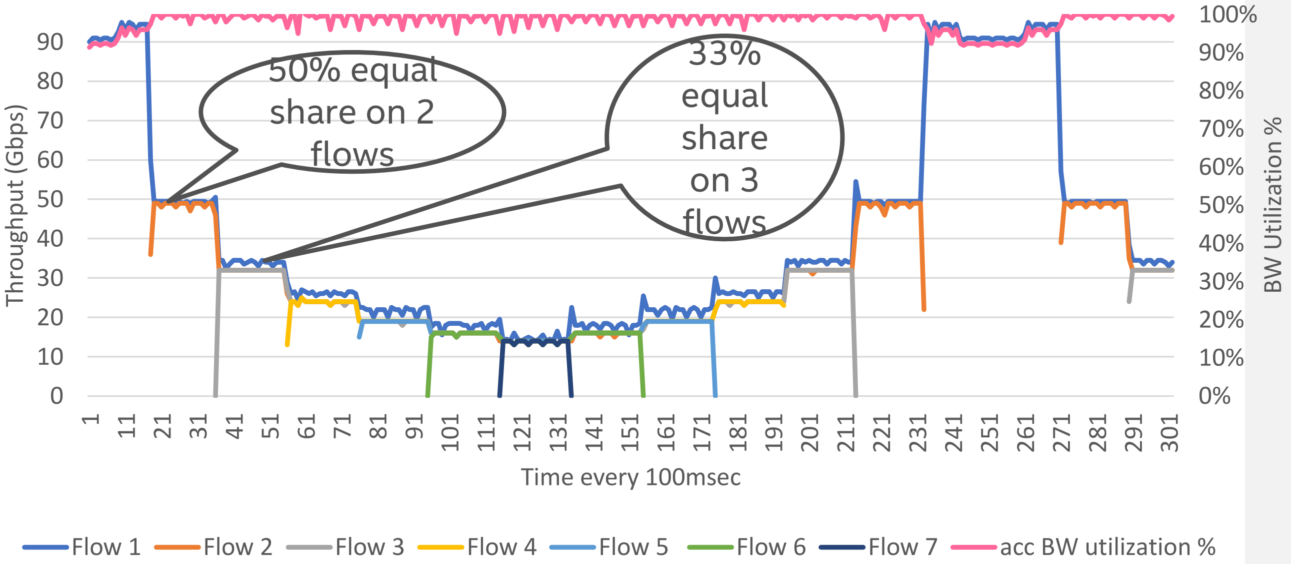


- PFC is great but does not work for multi-tenant and multi-level switches
- When packet drops occur, utilization is poor

SWIFT congestion control for Habana Gaudi2

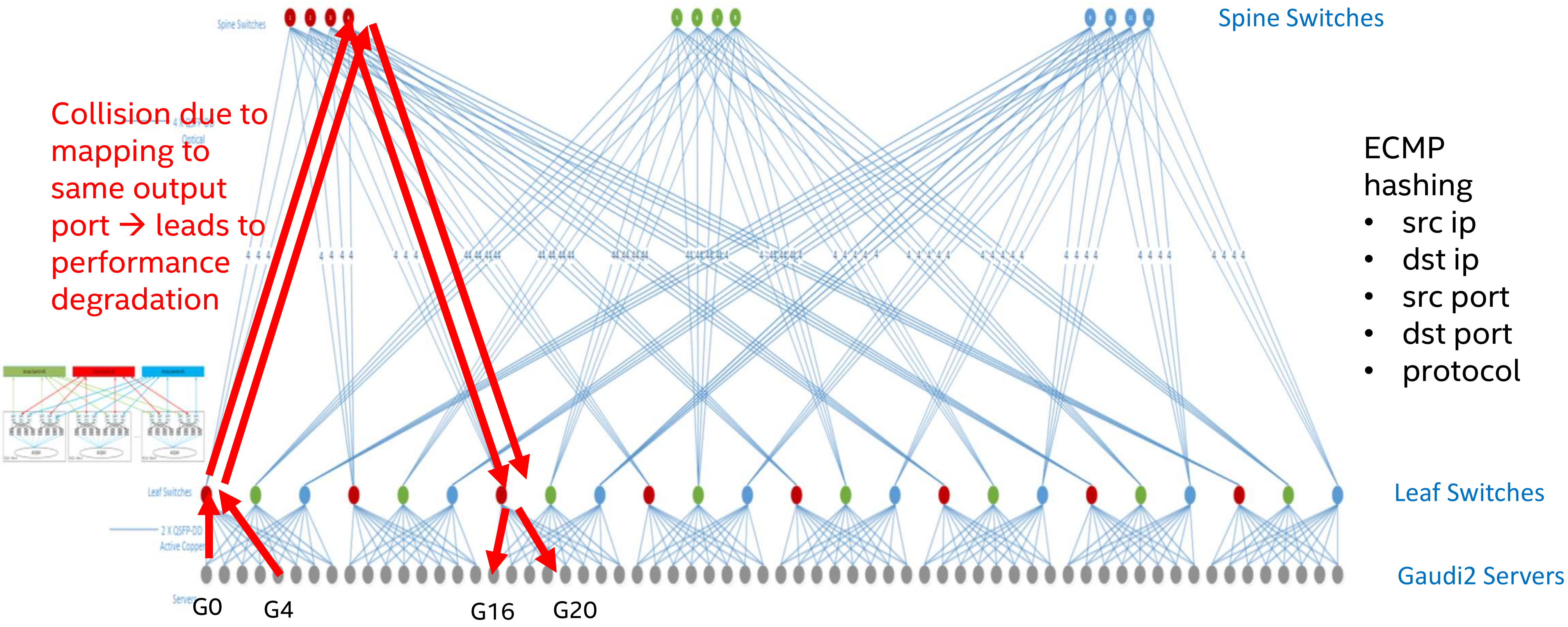
7:1 congestion, No PFC	#packets dropped	Bandwidth utilization
Default	276787	10-50%
SWIFT for 4KB/8KB MTU (targetdelay=20usecs, ai=2, beta=0.5,min_cwnd=2, max_cwnd=32)	~1	~98%





SWIFT:
https://dl.acm.org/doi/pdf/10.1145/3387514.3406591

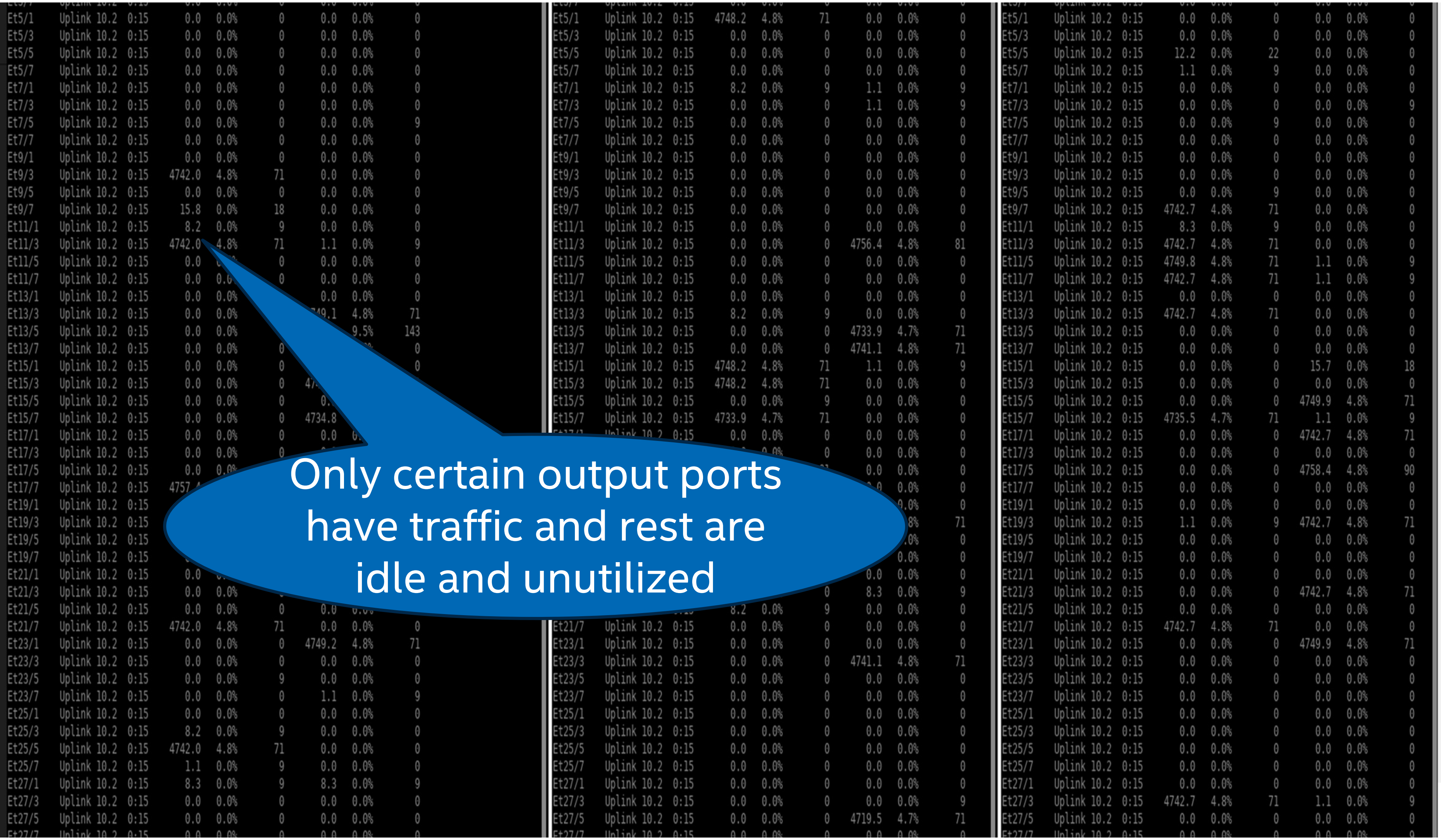
Packet collision at large-scale



Collision due to mapping to same output port → leads to performance degradation

ECMP hashing

- src ip
- dst ip
- src port
- dst port
- protocol



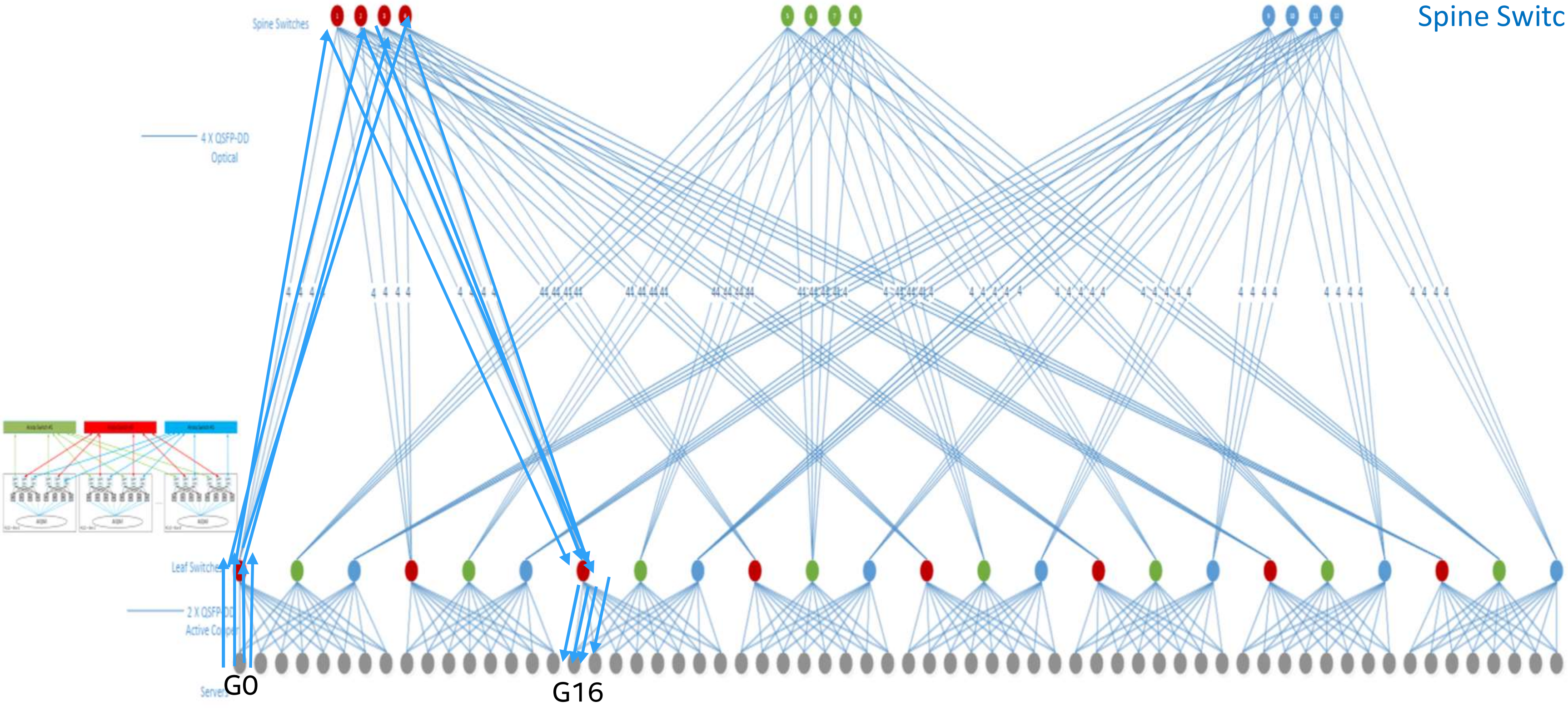
```
Et5/1   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et5/3   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et5/5   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et5/7   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et7/1   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et7/3   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et7/5   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     9
Et7/7   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/1   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/3   Uplink 10.2  0:15  4742.0  4.8%    71     0.0  0.0%     0
Et9/5   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/7   Uplink 10.2  0:15    15.8  0.0%    18     0.0  0.0%     0
Et11/1  Uplink 10.2  0:15     8.2  0.0%     9     0.0  0.0%     0
Et11/3  Uplink 10.2  0:15  4742.0  4.8%    71     1.1  0.0%     9
Et11/5  Uplink 10.2  0:15     0.0           0     0.0  0.0%     0
Et11/7  Uplink 10.2  0:15     0.0  0.0      0     0.0  0.0%     0
Et13/1  Uplink 10.2  0:15     0.0  0.0%           0.0  0.0%     0
Et13/3  Uplink 10.2  0:15     0.0  0.0%            49.1  4.8%    71
Et13/5  Uplink 10.2  0:15     0.0  0.0%                  9.5%   143
Et13/7  Uplink 10.2  0:15     0.0  0.0%     0                   0
Et15/1  Uplink 10.2  0:15     0.0  0.0%     0                   0
Et15/3  Uplink 10.2  0:15     0.0  0.0%     0   47
Et15/5  Uplink 10.2  0:15     0.0  0.0%     0     0.
Et15/7  Uplink 10.2  0:15     0.0  0.0%     0  4734.8
Et17/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.
Et17/3  Uplink 10.2  0:15     0.0  0.0%     0
Et17/5  Uplink 10.2  0:15     0.0  0.0%
Et17/7  Uplink 10.2  0:15  4757
Et19/1  Uplink 10.2  0:15
Et19/3  Uplink 10.2  0:15
Et19/5  Uplink 10.2  0:15
Et19/7  Uplink 10.2  0:15
Et21/1  Uplink 10.2  0:15     0.0
Et21/3  Uplink 10.2  0:15     0.0  0.0%
Et21/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0
Et21/7  Uplink 10.2  0:15  4742.0  4.8%    71     0.0  0.0%     0
Et23/1  Uplink 10.2  0:15     0.0  0.0%     0  4749.2  4.8%    71
Et23/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et23/5  Uplink 10.2  0:15     0.0  0.0%     9     0.0  0.0%     0
Et23/7  Uplink 10.2  0:15     0.0  0.0%     0     1.1  0.0%     9
Et25/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/3  Uplink 10.2  0:15     8.2  0.0%     9     0.0  0.0%     0
Et25/5  Uplink 10.2  0:15  4742.0  4.8%    71     0.0  0.0%     0
Et25/7  Uplink 10.2  0:15     1.1  0.0%     9     0.0  0.0%     0
Et27/1  Uplink 10.2  0:15     8.3  0.0%     9     8.3  0.0%     9
Et27/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et27/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et27/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
```

```
Et5/1   Uplink 10.2  0:15  4748.2  4.8%    71     0.0  0.0%     0
Et5/3   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et5/5   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et5/7   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et7/1   Uplink 10.2  0:15     8.2  0.0%     9     1.1  0.0%     9
Et7/3   Uplink 10.2  0:15     0.0  0.0%     0     1.1  0.0%     9
Et7/5   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et7/7   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/1   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/3   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/5   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/7   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et11/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et11/3  Uplink 10.2  0:15     0.0  0.0%     0  4756.4  4.8%    81
Et11/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et11/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et13/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et13/3  Uplink 10.2  0:15     8.2  0.0%     9     0.0  0.0%     0
Et13/5  Uplink 10.2  0:15     0.0  0.0%     0  4733.9  4.7%    71
Et13/7  Uplink 10.2  0:15     0.0  0.0%     0  4741.1  4.8%    71
Et15/1  Uplink 10.2  0:15  4748.2  4.8%    71     1.1  0.0%     9
Et15/3  Uplink 10.2  0:15  4748.2  4.8%    71     0.0  0.0%     0
Et15/5  Uplink 10.2  0:15     0.0  0.0%     9     0.0  0.0%     0
Et15/7  Uplink 10.2  0:15  4733.9  4.7%    71     0.0  0.0%     0
Et17/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
                                   0.0%     0     0.0  0.0%     0
                                                  0.0  0.0%     0
                                                  0.0  0.0%     0
                                                        .0%     0
                                                         8%    71
                                                        .0%     0
                                                       0.0%     0
                                                  0.0  0.0%     0
                                            0     8.3  0.0%     9
                              8.2  0.0%     9     0.0  0.0%     0
Et21/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et23/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et23/3  Uplink 10.2  0:15     0.0  0.0%     0  4741.1  4.8%    71
Et23/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et23/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et27/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et27/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     9
Et27/5  Uplink 10.2  0:15     0.0  0.0%     0  4719.5  4.7%    71
Et27/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
```

```
Et5/1   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et5/3   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et5/5   Uplink 10.2  0:15    12.2  0.0%    22     0.0  0.0%     0
Et5/7   Uplink 10.2  0:15     1.1  0.0%     9     0.0  0.0%     0
Et7/1   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et7/3   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     9
Et7/5   Uplink 10.2  0:15     0.0  0.0%     9     0.0  0.0%     0
Et7/7   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/1   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/3   Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et9/5   Uplink 10.2  0:15     0.0  0.0%     9     0.0  0.0%     0
Et9/7   Uplink 10.2  0:15  4742.7  4.8%    71     0.0  0.0%     0
Et11/1  Uplink 10.2  0:15     8.3  0.0%     9     0.0  0.0%     0
Et11/3  Uplink 10.2  0:15  4742.7  4.8%    71     0.0  0.0%     0
Et11/5  Uplink 10.2  0:15  4749.8  4.8%    71     1.1  0.0%     9
Et11/7  Uplink 10.2  0:15  4742.7  4.8%    71     1.1  0.0%     9
Et13/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et13/3  Uplink 10.2  0:15  4742.7  4.8%    71     0.0  0.0%     0
Et13/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et13/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et15/1  Uplink 10.2  0:15     0.0  0.0%     0    15.7  0.0%    18
Et15/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et15/5  Uplink 10.2  0:15     0.0  0.0%     0  4749.9  4.8%    71
Et15/7  Uplink 10.2  0:15  4735.5  4.7%    71     1.1  0.0%     9
Et17/1  Uplink 10.2  0:15     0.0  0.0%     0  4742.7  4.8%    71
Et17/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et17/5  Uplink 10.2  0:15     0.0  0.0%     0  4758.4  4.8%    90
Et17/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et19/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et19/3  Uplink 10.2  0:15     1.1  0.0%     9  4742.7  4.8%    71
Et19/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et19/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et21/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et21/3  Uplink 10.2  0:15     0.0  0.0%     0  4742.7  4.8%    71
Et21/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et21/7  Uplink 10.2  0:15  4742.7  4.8%    71     0.0  0.0%     0
Et23/1  Uplink 10.2  0:15     0.0  0.0%     0  4749.9  4.8%    71
Et23/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et23/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et23/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/3  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et25/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et27/1  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et27/3  Uplink 10.2  0:15  4742.7  4.8%    71     1.1  0.0%     9
Et27/5  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     0
Et27/7  Uplink 10.2  0:15     0.0  0.0%     0     0.0  0.0%     9
```

Only certain output ports have traffic and rest are idle and unutilized

Solution: Packet spraying

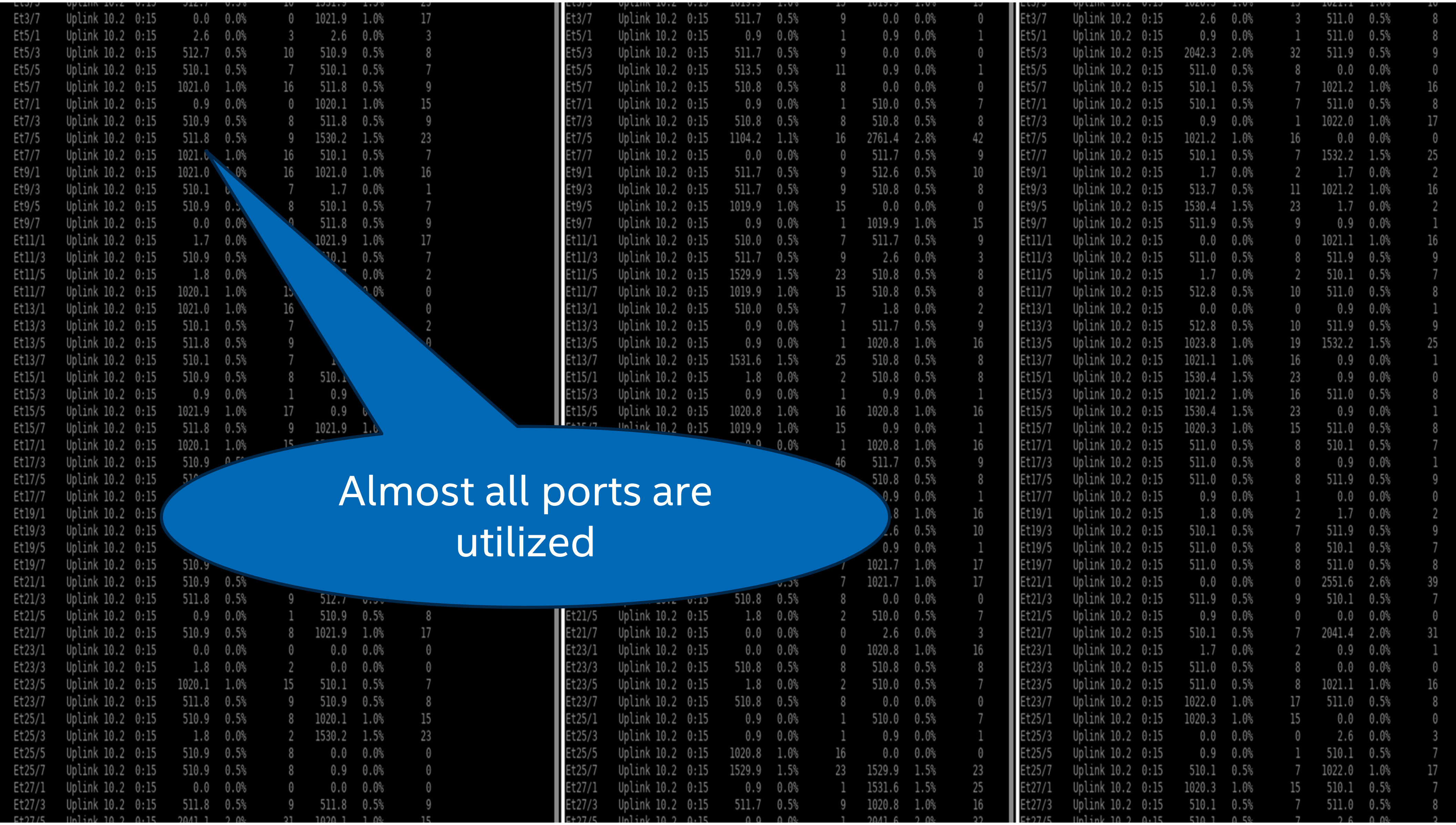


ECMP hashing

- src ip
- dst ip
- src port
- dst port
- protocol

Packet spraying solution

HCCL collectives BW	With collisions	Packet spraying	Expected
All2All	22 GB/s	64 GB/s	65 GB/s
Allgather	183 GB/s	272 GB/s	272 GB/s



```
Et3/7     Uplink 10.2  0:15     0.0  0.0%      0  1021.9  1.0%     17
Et5/1     Uplink 10.2  0:15     2.6  0.0%      3     2.6  0.0%      3
Et5/3     Uplink 10.2  0:15   512.7  0.5%     10   510.9  0.5%      8
Et5/5     Uplink 10.2  0:15   510.1  0.5%      7   510.1  0.5%      7
Et5/7     Uplink 10.2  0:15  1021.0  1.0%     16   511.8  0.5%      9
Et7/1     Uplink 10.2  0:15     0.9  0.0%      0  1020.1  1.0%     15
Et7/3     Uplink 10.2  0:15   510.9  0.5%      8   511.8  0.5%      9
Et7/5     Uplink 10.2  0:15   511.8  0.5%      9  1530.2  1.5%     23
Et7/7     Uplink 10.2  0:15  1021.0  1.0%     16   510.1  0.5%      7
Et9/1     Uplink 10.2  0:15  1021.0            16  1021.0  1.0%     16
Et9/3     Uplink 10.2  0:15   510.1             7     1.7  0.0%      1
Et9/5     Uplink 10.2  0:15   510.9             8   510.1  0.5%      7
Et9/7     Uplink 10.2  0:15     0.0  0.0%          511.8  0.5%      9
Et11/1    Uplink 10.2  0:15     1.7  0.0%         1021.9  1.0%     17
Et11/3    Uplink 10.2  0:15   510.9  0.5%                 0.5%      7
Et11/5    Uplink 10.2  0:15     1.8  0.0%                 0.0%      2
Et11/7    Uplink 10.2  0:15  1020.1  1.0%                           0
Et13/1    Uplink 10.2  0:15  1021.0  1.0%     16                    0
Et13/3    Uplink 10.2  0:15   510.1  0.5%      7                    2
Et13/5    Uplink 10.2  0:15   511.8  0.5%      9                    0
Et13/7    Uplink 10.2  0:15   510.1  0.5%      7
Et15/1    Uplink 10.2  0:15   510.9  0.5%      8
Et15/3    Uplink 10.2  0:15     0.9  0.0%      1     0.9
Et15/5    Uplink 10.2  0:15  1021.9  1.0%     17     0.9
Et15/7    Uplink 10.2  0:15   511.8  0.5%      9  1021.9
Et17/1    Uplink 10.2  0:15  1020.1  1.0%
Et17/3    Uplink 10.2  0:15   510.9
Et17/5    Uplink 10.2  0:15
Et17/7    Uplink 10.2  0:15
Et19/1    Uplink 10.2  0:15
Et19/3    Uplink 10.2  0:15
Et19/5    Uplink 10.2  0:15
Et19/7    Uplink 10.2  0:15   510.9
Et21/1    Uplink 10.2  0:15   510.9  0.5%
Et21/3    Uplink 10.2  0:15   511.8  0.5%      9
Et21/5    Uplink 10.2  0:15     0.9  0.0%      1   510.9  0.5%      8
Et21/7    Uplink 10.2  0:15   510.9  0.5%      8  1021.9  1.0%     17
Et23/1    Uplink 10.2  0:15     0.0  0.0%      0     0.0  0.0%      0
Et23/3    Uplink 10.2  0:15     1.8  0.0%      2     0.0  0.0%      0
Et23/5    Uplink 10.2  0:15  1020.1  1.0%     15   510.1  0.5%      7
Et23/7    Uplink 10.2  0:15   511.8  0.5%      9   510.9  0.5%      8
Et25/1    Uplink 10.2  0:15   510.9  0.5%      8  1020.1  1.0%     15
Et25/3    Uplink 10.2  0:15     1.8  0.0%      2  1530.2  1.5%     23
Et25/5    Uplink 10.2  0:15   510.9  0.5%      8     0.0  0.0%      0
Et25/7    Uplink 10.2  0:15   510.9  0.5%      8     0.9  0.0%      0
Et27/1    Uplink 10.2  0:15     0.0  0.0%      0     0.0  0.0%      0
Et27/3    Uplink 10.2  0:15   511.8  0.5%      9   511.8  0.5%      9
Et27/5    Uplink 10.2  0:15  2041.1  2.0%     31  1020.1  1.0%     15
```

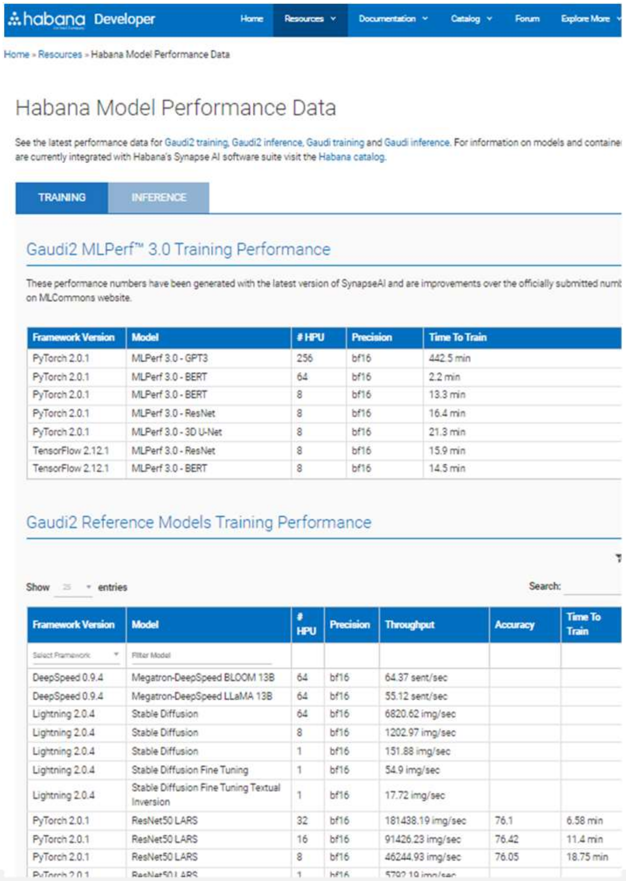
```
Et3/7     Uplink 10.2  0:15   511.7  0.5%      9     0.0  0.0%      0
Et5/1     Uplink 10.2  0:15     0.9  0.0%      1     0.9  0.0%      1
Et5/3     Uplink 10.2  0:15   511.7  0.5%      9     0.0  0.0%      0
Et5/5     Uplink 10.2  0:15   513.5  0.5%     11     0.9  0.0%      1
Et5/7     Uplink 10.2  0:15   510.8  0.5%      8     0.0  0.0%      0
Et7/1     Uplink 10.2  0:15     0.9  0.0%      1   510.0  0.5%      7
Et7/3     Uplink 10.2  0:15   510.8  0.5%      8   510.8  0.5%      8
Et7/5     Uplink 10.2  0:15  1104.2  1.1%     16  2761.4  2.8%     42
Et7/7     Uplink 10.2  0:15     0.0  0.0%      0   511.7  0.5%      9
Et9/1     Uplink 10.2  0:15   511.7  0.5%      9   512.6  0.5%     10
Et9/3     Uplink 10.2  0:15   511.7  0.5%      9   510.8  0.5%      8
Et9/5     Uplink 10.2  0:15  1019.9  1.0%     15     0.0  0.0%      0
Et9/7     Uplink 10.2  0:15     0.9  0.0%      1  1019.9  1.0%     15
Et11/1    Uplink 10.2  0:15   510.0  0.5%      7   511.7  0.5%      9
Et11/3    Uplink 10.2  0:15   511.7  0.5%      9     2.6  0.0%      3
Et11/5    Uplink 10.2  0:15  1529.9  1.5%     23   510.8  0.5%      8
Et11/7    Uplink 10.2  0:15  1019.9  1.0%     15   510.8  0.5%      8
Et13/1    Uplink 10.2  0:15   510.0  0.5%      7     1.8  0.0%      2
Et13/3    Uplink 10.2  0:15     0.9  0.0%      1   511.7  0.5%      9
Et13/5    Uplink 10.2  0:15     0.9  0.0%      1  1020.8  1.0%     16
Et13/7    Uplink 10.2  0:15  1531.6  1.5%     25   510.8  0.5%      8
Et15/1    Uplink 10.2  0:15     1.8  0.0%      2   510.8  0.5%      8
Et15/3    Uplink 10.2  0:15     0.9  0.0%      1     0.9  0.0%      1
Et15/5    Uplink 10.2  0:15  1020.8  1.0%     16  1020.8  1.0%     16
          Uplink 10.2  0:15  1019.9  1.0%     15     0.9  0.0%      1
                                0.9  0.0%      1  1020.8  1.0%     16
                                              46   511.7  0.5%      9
                                                   510.8  0.5%      8
                                                     0.9  0.0%      1
                                                          1.0%     16
                                                          0.5%     10
                                                     0.9  0.0%      1
                                               7  1021.7  1.0%     17
                                               7  1021.7  1.0%     17
                              510.8  0.5%      8     0.0  0.0%      0
Et21/5    Uplink 10.2  0:15     1.8  0.0%      2   510.0  0.5%      7
Et21/7    Uplink 10.2  0:15     0.0  0.0%      0     2.6  0.0%      3
Et23/1    Uplink 10.2  0:15     0.0  0.0%      0  1020.8  1.0%     16
Et23/3    Uplink 10.2  0:15   510.8  0.5%      8   510.8  0.5%      8
Et23/5    Uplink 10.2  0:15     1.8  0.0%      2   510.0  0.5%      7
Et23/7    Uplink 10.2  0:15   510.8  0.5%      8     0.0  0.0%      0
Et25/1    Uplink 10.2  0:15     0.9  0.0%      1   510.0  0.5%      7
Et25/3    Uplink 10.2  0:15     0.9  0.0%      1     0.9  0.0%      1
Et25/5    Uplink 10.2  0:15  1020.8  1.0%     16     0.0  0.0%      0
Et25/7    Uplink 10.2  0:15  1529.9  1.5%     23  1529.9  1.5%     23
Et27/1    Uplink 10.2  0:15     0.9  0.0%      1  1531.6  1.5%     25
Et27/3    Uplink 10.2  0:15   511.7  0.5%      9  1020.8  1.0%     16
Et27/5    Uplink 10.2  0:15     0.9  0.0%      1  2041.6  2.0%     32
```

```
Et3/7     Uplink 10.2  0:15     2.6  0.0%      3   511.0  0.5%      8
Et5/1     Uplink 10.2  0:15     0.9  0.0%      1   511.0  0.5%      8
Et5/3     Uplink 10.2  0:15  2042.3  2.0%     32   511.9  0.5%      9
Et5/5     Uplink 10.2  0:15   511.0  0.5%      8     0.0  0.0%      0
Et5/7     Uplink 10.2  0:15   510.1  0.5%      7  1021.2  1.0%     16
Et7/1     Uplink 10.2  0:15   510.1  0.5%      7   511.0  0.5%      8
Et7/3     Uplink 10.2  0:15     0.9  0.0%      1  1022.0  1.0%     17
Et7/5     Uplink 10.2  0:15  1021.2  1.0%     16     0.0  0.0%      0
Et7/7     Uplink 10.2  0:15   510.1  0.5%      7  1532.2  1.5%     25
Et9/1     Uplink 10.2  0:15     1.7  0.0%      2     1.7  0.0%      2
Et9/3     Uplink 10.2  0:15   513.7  0.5%     11  1021.2  1.0%     16
Et9/5     Uplink 10.2  0:15  1530.4  1.5%     23     1.7  0.0%      2
Et9/7     Uplink 10.2  0:15   511.9  0.5%      9     0.9  0.0%      1
Et11/1    Uplink 10.2  0:15     0.0  0.0%      0  1021.1  1.0%     16
Et11/3    Uplink 10.2  0:15   511.0  0.5%      8   511.9  0.5%      9
Et11/5    Uplink 10.2  0:15     1.7  0.0%      2   510.1  0.5%      7
Et11/7    Uplink 10.2  0:15   512.8  0.5%     10   511.0  0.5%      8
Et13/1    Uplink 10.2  0:15     0.0  0.0%      0     0.9  0.0%      1
Et13/3    Uplink 10.2  0:15   512.8  0.5%     10   511.9  0.5%      9
Et13/5    Uplink 10.2  0:15  1023.8  1.0%     19  1532.2  1.5%     25
Et13/7    Uplink 10.2  0:15  1021.1  1.0%     16     0.9  0.0%      1
Et15/1    Uplink 10.2  0:15  1530.4  1.5%     23     0.9  0.0%      0
Et15/3    Uplink 10.2  0:15  1021.2  1.0%     16   511.0  0.5%      8
Et15/5    Uplink 10.2  0:15  1530.4  1.5%     23     0.9  0.0%      1
Et15/7    Uplink 10.2  0:15  1020.3  1.0%     15   511.0  0.5%      8
Et17/1    Uplink 10.2  0:15   511.0  0.5%      8   510.1  0.5%      7
Et17/3    Uplink 10.2  0:15   511.0  0.5%      8     0.9  0.0%      1
Et17/5    Uplink 10.2  0:15   511.0  0.5%      8   511.9  0.5%      9
Et17/7    Uplink 10.2  0:15     0.9  0.0%      1     0.0  0.0%      0
Et19/1    Uplink 10.2  0:15     1.8  0.0%      2     1.7  0.0%      2
Et19/3    Uplink 10.2  0:15   510.1  0.5%      7   511.9  0.5%      9
Et19/5    Uplink 10.2  0:15   511.0  0.5%      8   510.1  0.5%      7
Et19/7    Uplink 10.2  0:15   511.0  0.5%      8   511.0  0.5%      8
Et21/1    Uplink 10.2  0:15     0.0  0.0%      0  2551.6  2.6%     39
Et21/3    Uplink 10.2  0:15   511.9  0.5%      9   510.1  0.5%      7
Et21/5    Uplink 10.2  0:15     0.9  0.0%      0     0.0  0.0%      0
Et21/7    Uplink 10.2  0:15   510.1  0.5%      7  2041.4  2.0%     31
Et23/1    Uplink 10.2  0:15     1.7  0.0%      2     0.9  0.0%      1
Et23/3    Uplink 10.2  0:15   511.0  0.5%      8     0.0  0.0%      0
Et23/5    Uplink 10.2  0:15   511.0  0.5%      8  1021.1  1.0%     16
Et23/7    Uplink 10.2  0:15  1022.0  1.0%     17   511.0  0.5%      8
Et25/1    Uplink 10.2  0:15  1020.3  1.0%     15     0.0  0.0%      0
Et25/3    Uplink 10.2  0:15     0.0  0.0%      0     2.6  0.0%      3
Et25/5    Uplink 10.2  0:15     0.9  0.0%      1   510.1  0.5%      7
Et25/7    Uplink 10.2  0:15   510.1  0.5%      7  1022.0  1.0%     17
Et27/1    Uplink 10.2  0:15  1020.3  1.0%     15   510.1  0.5%      7
Et27/3    Uplink 10.2  0:15   510.1  0.5%      7   511.0  0.5%      8
Et27/5    Uplink 10.2  0:15   510.1  0.5%      7     2.6  0.0%      3
```

Almost all ports are utilized

Developer Resources

Gaudi Developer Site: developer.Habana.ai



Habana Model Performance Data

See the latest performance data for Gaudi2 training, Gaudi2 inference, Gaudi training and Gaudi inference. For information on models and containe are currently integrated with Habana's Synapse AI software suite visit the Habana catalog.

Gaudi2 MLPerf™ 3.0 Training Performance

These performance numbers have been generated with the latest version of SynapseAI and are improvements over the officially submitted numb on MLCommons website.

Framework Version	Model	# HPU	Precision	Time To Train
PyTorch 2.0.1	MLPerf 3.0 - GPT3	256	bf16	442.5 min
PyTorch 2.0.1	MLPerf 3.0 - BERT	64	bf16	2.2 min
PyTorch 2.0.1	MLPerf 3.0 - BERT	8	bf16	13.3 min
PyTorch 2.0.1	MLPerf 3.0 - ResNet	8	bf16	16.4 min
PyTorch 2.0.1	MLPerf 3.0 - 3D U-Net	8	bf16	21.3 min
TensorFlow 2.12.1	MLPerf 3.0 - ResNet	8	bf16	15.9 min
TensorFlow 2.12.1	MLPerf 3.0 - BERT	8	bf16	14.5 min

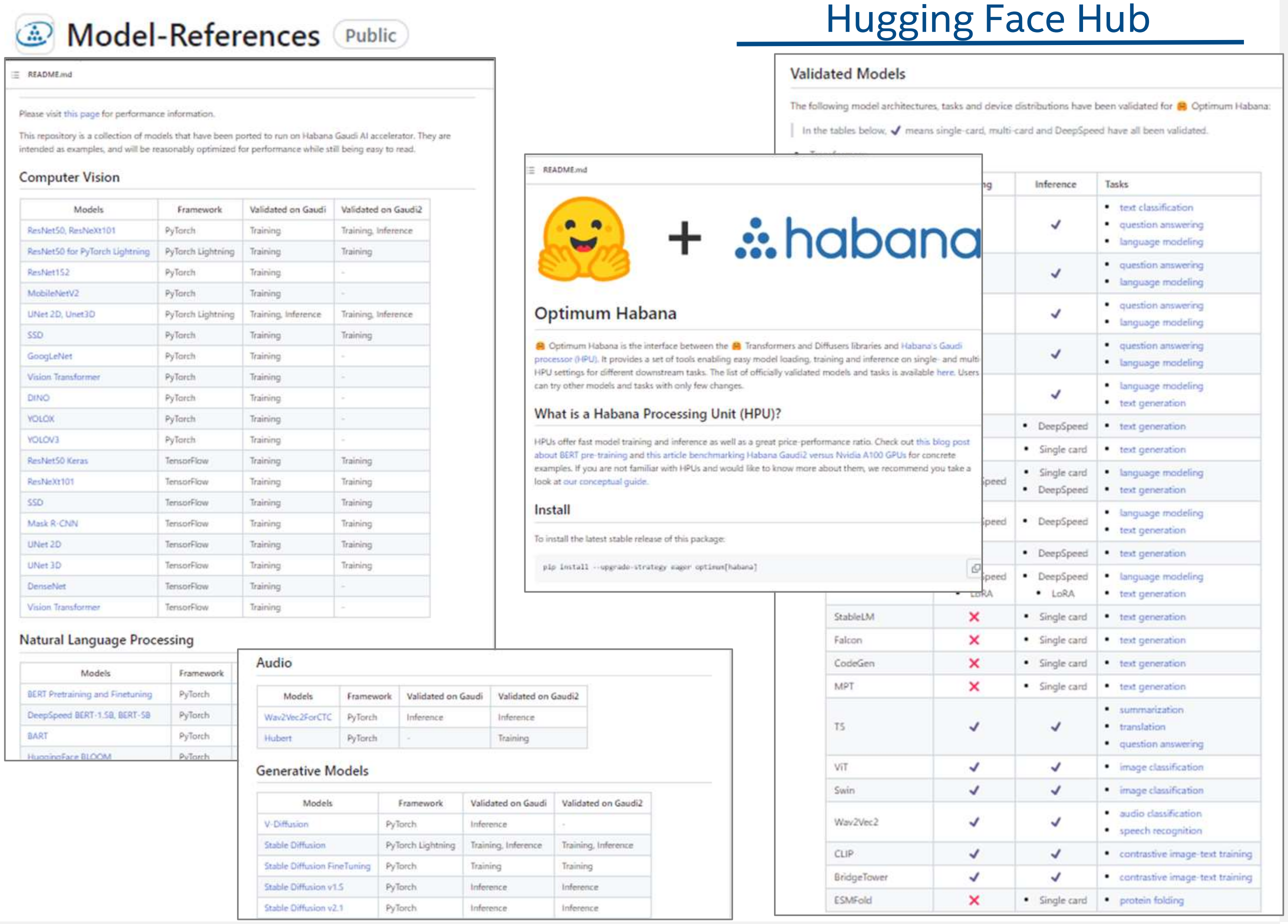
Gaudi2 Reference Models Training Performance

Show 25 entries Search:

Framework Version	Model	# HPU	Precision	Throughput	Accuracy	Time To Train
Select Framework	Filter Model					
DeepSpeed 0.9.4	Megatron-DeepSpeed BLOOM 13B	64	bf16	64.37 sent/sec		
DeepSpeed 0.9.4	Megatron-DeepSpeed LLaMA 13B	64	bf16	55.12 sent/sec		
Lightning 2.0.4	Stable Diffusion	64	bf16	6820.62 img/sec		
Lightning 2.0.4	Stable Diffusion	8	bf16	1202.97 img/sec		
Lightning 2.0.4	Stable Diffusion	1	bf16	151.88 img/sec		
Lightning 2.0.4	Stable Diffusion Fine Tuning	1	bf16	54.9 img/sec		
Lightning 2.0.4	Stable Diffusion Fine Tuning Textual Inversion	1	bf16	17.72 img/sec		
PyTorch 2.0.1	ResNet50 LARS	32	bf16	181438.19 img/sec	76.1	6.58 min
PyTorch 2.0.1	ResNet50 LARS	16	bf16	91426.23 img/sec	76.42	11.4 min
PyTorch 2.0.1	ResNet50 LARS	8	bf16	46244.93 img/sec	76.05	18.75 min
PyTorch 2.0.1	ResNet50 LARS	1	bf16	5792.19 img/sec		

Habana GitHub

Habana Optimum Library on Hugging Face Hub



Please visit this page for performance information.

This repository is a collection of models that have been ported to run on Habana Gaudi AI accelerator. They are intended as examples, and will be reasonably optimized for performance while still being easy to read.

Computer Vision

Models	Framework	Validated on Gaudi	Validated on Gaudi2
ResNet50, ResNeXt101	PyTorch	Training	Training, Inference
ResNet50 for PyTorch Lightning	PyTorch Lightning	Training	Training
ResNet152	PyTorch	Training	-
MobileNetV2	PyTorch	Training	-
UNet 2D, Unet3D	PyTorch Lightning	Training, Inference	Training, Inference
SSD	PyTorch	Training	Training
GoogLeNet	PyTorch	Training	-
Vision Transformer	PyTorch	Training	-
DINO	PyTorch	Training	-
YOLOX	PyTorch	Training	-
YOLOV3	PyTorch	Training	-
ResNet50 Keras	TensorFlow	Training	Training
ResNeXt101	TensorFlow	Training	Training
SSD	TensorFlow	Training	Training
Mask R-CNN	TensorFlow	Training	Training
UNet 2D	TensorFlow	Training	Training
UNet 3D	TensorFlow	Training	Training
DenseNet	TensorFlow	Training	-
Vision Transformer	TensorFlow	Training	-

Natural Language Processing

Models	Framework
BERT Pretraining and Finetuning	PyTorch
DeepSpeed BERT-1.5B, BERT-5B	PyTorch
BART	PyTorch
HuggingFace BLOOM	PyTorch

Audio

Models	Framework	Validated on Gaudi	Validated on Gaudi2
Wav2Vec2ForCTC	PyTorch	Inference	Inference
Hubert	PyTorch	-	Training

Generative Models

Models	Framework	Validated on Gaudi	Validated on Gaudi2
V-Diffusion	PyTorch	Inference	-
Stable Diffusion	PyTorch Lightning	Training, Inference	Training, Inference
Stable Diffusion FineTuning	PyTorch	Training	Training
Stable Diffusion v1.5	PyTorch	Inference	Inference
Stable Diffusion v2.1	PyTorch	Inference	Inference

Optimum Habana

🤗 Optimum Habana is the interface between the 🤗 Transformers and Diffusers libraries and Habana's Gaudi processor (HPU). It provides a set of tools enabling easy model loading, training and inference on single- and multi-HPU settings for different downstream tasks. The list of officially validated models and tasks is available here. Users can try other models and tasks with only few changes.

What is a Habana Processing Unit (HPU)?

HPUs offer fast model training and inference as well as a great price-performance ratio. Check out this blog post about BERT pre-training and this article benchmarking Habana Gaudi2 versus Nvidia A100 GPUs for concrete examples. If you are not familiar with HPUs and would like to know more about them, we recommend you take a look at our conceptual guide.

Install

To install the latest stable release of this package:

```
pip install --upgrade-strategy eager optimum[habana]
```

Validated Models

The following model architectures, tasks and device distributions have been validated for 🤗 Optimum Habana:

> In the tables below, ✓ means single-card, multi-card and DeepSpeed have all been validated.

	ng	Inference	Tasks
		✓	• text classification • question answering • language modeling
		✓	• question answering • language modeling
		✓	• question answering • language modeling
		✓	• question answering • language modeling
		✓	• language modeling • text generation
		• DeepSpeed	• text generation
		• Single card	• text generation
	peed	• Single card • DeepSpeed	• language modeling • text generation
	peed	• DeepSpeed	• language modeling • text generation
		• DeepSpeed	• text generation
	peed • LoRA	• DeepSpeed • LoRA	• language modeling • text generation
StableLM	✗	• Single card	• text generation
Falcon	✗	• Single card	• text generation
CodeGen	✗	• Single card	• text generation
MPT	✗	• Single card	• text generation
T5	✓	✓	• summarization • translation • question answering
ViT	✓	✓	• image classification
Swin	✓	✓	• image classification
Wav2Vec2	✓	✓	• audio classification • speech recognition
CLIP	✓	✓	• contrastive image-text training
BridgeTower	✓	✓	• contrastive image-text training
ESMFold	✗	• Single card	• protein folding

Summary

- Intel Gaudi2 continues to be the only viable alternative to NVIDIA's H100 for GenAI/LLM compute, with a significant price-performance advantage.
- 4th Gen Intel Xeon processors help customers train small- to mid-sized deep learning models, as well as fine tuning and transfer learning.
- Intel is well positioned to address every phase of the AI continuum across AI workloads, from large to small models—giving customers choice.

Thank you