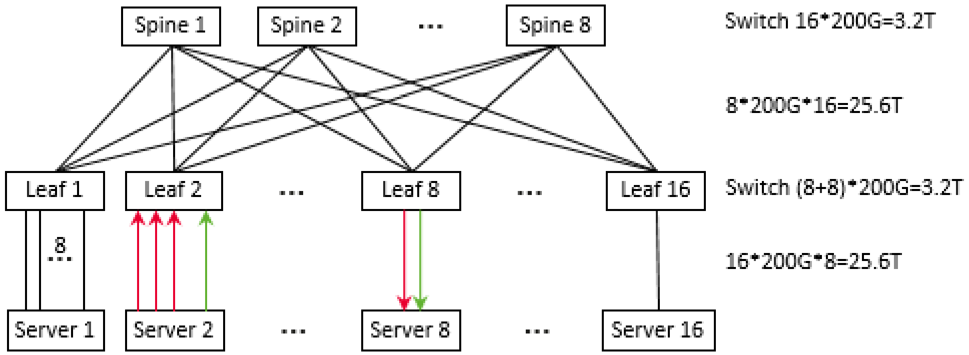
SFC simulation

Lihao Chen (lihao.chen@huawei.com)
Sivakolundu, Ramesh (sramesh@cisco.com)
Paul Congdon (paul.congdon@outlook.com)
Lily Lv (lvyunping@huawei.com)

Xiaowu (Horace) He (hexiaowu@huawei.com) - Simulator development
Siyu Yan (yansiyu@huawei.com) - AI use case configuration

Background & Introduction

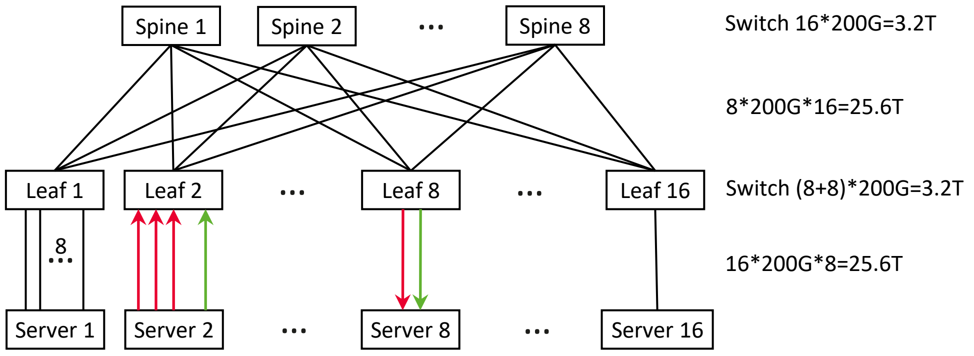
- Disposition of https://www.ieee802.org/1/files/public/docs2025/dw-chen-sfc-consideration-and-design-0725-v01.pdf
  - > Discussion summary: Collaborators to submit SFC use cases for lossless and best-effort networks, and provide comparative performance data against existing methods (PFC, ECN+QCN).
- This presentation shows simulation results of SFC in AIDC backend network use cases.
  - > 2-layer CLOS topology, one with 128 GPUs, another with 2048 GPUs. Collective communication flow characteristics.
  - > Detailed settings see next pages.
- Simulator: HTSIM (This platform is used by UEC for simulation works).
  - > The SFC function is newly developed. Class-based, pause mode.
  - > We didn't use OQ (output queue) model, as it would literally spread PFC 'everywhere' during congestion, making it too easy for SFC to outperform PFC.
  - > We use the IQ (input queue) model with modifications, resolved the issue that victim flows can be wrongly SFCed because of the HoLB of IQ.
  - > HTSIM currently lacks support for VOQ (Virtual Output Queuing), but its behavior now resembles VOQ in the context of PFC and SFC operation. VOQ remains the primary queuing architecture in data center switches.



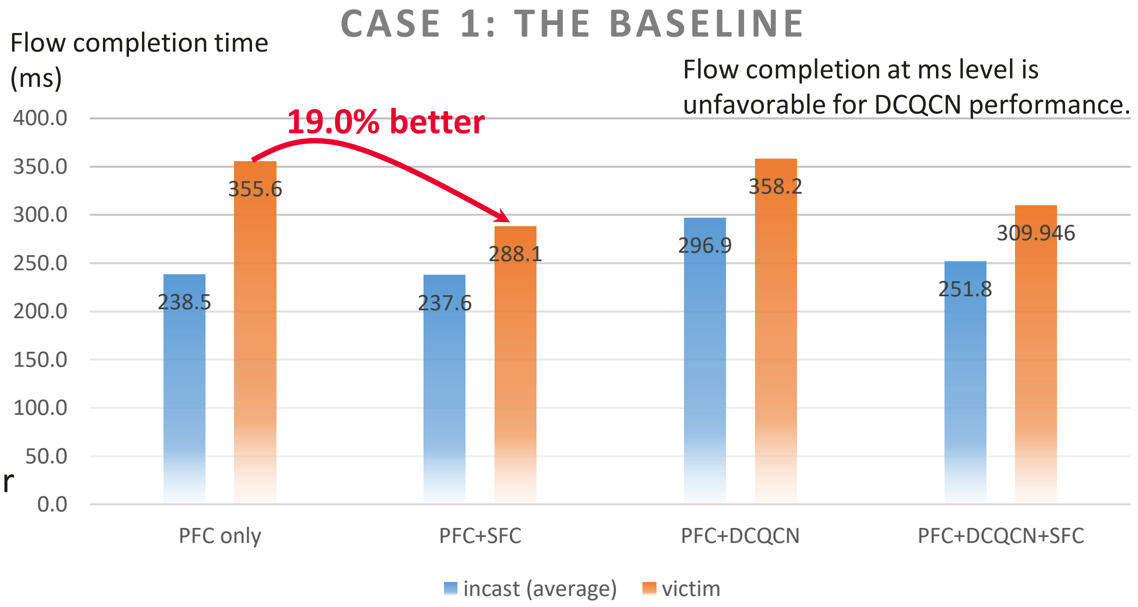
8 GPU per server, 128 GPU in total

Simulation Part 1: prove of concept

- Network 2-layer CLOS
  - > 8 spine switches, 16*200G. 16 Leaf switches, 16*200G
  - > 128 GPU, 200G
- Basic settings
  - > Link delay: 150ns. Switch process (fixed) delay: 300ns
  - > Max. packet size: 4KB
  - > Buffer size: 400KB. PFC threshold: 380KB (when there is only 20KB buffer left, PFC kicks in). DCQCN K: 200KB.
- Flow settings
  - > 3 flows to 1 incast (red arrows), 2MB data per flow
  - > The other flow (the victim, green arrow), 5MB data
- User-defined SFC settings:
  - > SFC threshold: 200KB; SFC pause time: 6us; SFCM Min. interval: 6us.
  - > *Based on https://www.ieee802.org/1/files/public/docs2025/dw-chen-sfc-computation-simulation.pdf , SFC headroom should be larger than 120KB and SFC pause time should be within (4.8, 12.8) µs.



8 GPU per server, 128 GPU in total



19.0% better

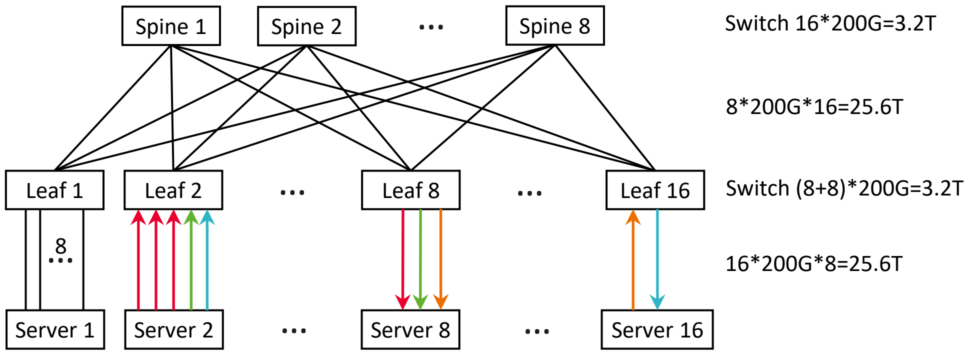
Flow completion at ms level is unfavorable for DCQCN performance.

SFC precisely pauses the source of incasts, improving the performance of the victim flow.

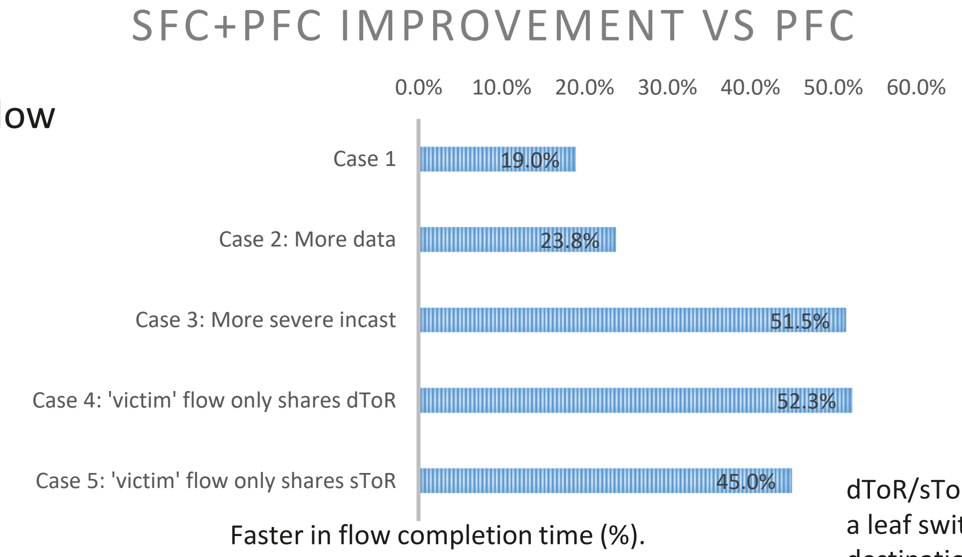
Simulation Part 1: prove of concept Cont.

- Case 2: Change flow settings from Case 1
  > 3-to-1 incast, 2MB -> 20MB data; The other flow, 5MB -> 50MB data
- Case 3: Change flow settings from Case 1
  > 3-to-1 incast -> 6-to-1 incast
  > Adjust SFC settings accordingly
    - SFC threshold: 80KB; SFC pause time: 15us
    - SFCM minimum interval: 15us
- Case 4 & 5: Based on Case 3, observe the orange / blue flow

Flow completion time (ms)		PFC only	PFC+SFC
Case 1	incast	238.5053	237.6383
	victim	355.553	288.14
Case 2	incast	2398.043	2355.47
	victim	3738.1	2848.46
Case 3	incast	418.9202	449.4542
	victim	533.231	258.392
Case 4	incast	426.9668	449.5625
	victim	543.147	258.88
Case 5	incast	417.2132	449.4542
	victim	470.107	258.392



8 GPU per server, 128 GPU in total

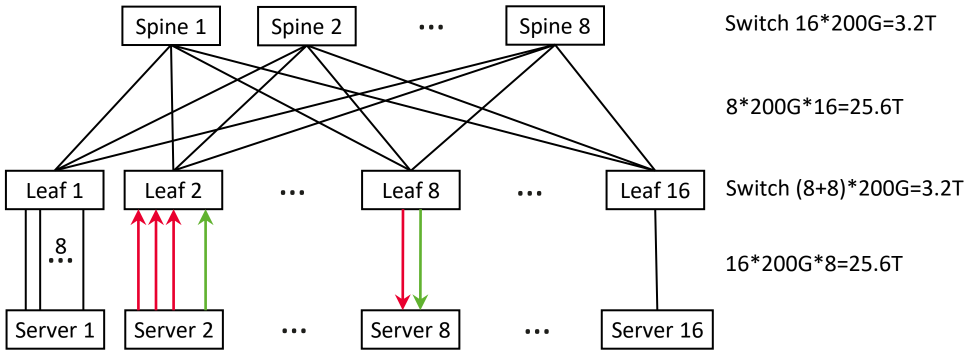


dToR/sToR: destination/source Top-of-Rack, a leaf switch that connects to the destination/source of the flow.

SFC improvements varies in different scenarios.

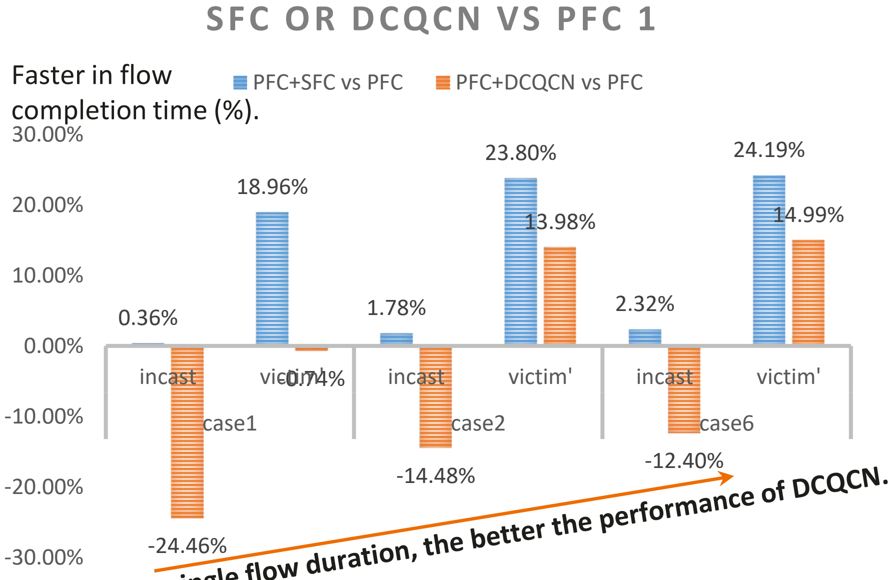
Simulation Part 1: prove of concept Cont.

- Case 2: Change flow settings based on Case 1
  > 3-to-1 incast, 2MB -> 20MB data; The other flow, 5MB -> 50MB data
- Case 6: Change flow settings based on Case 1
  > 3-to-1 incast, 2MB -> 200MB data; The other flow, 5MB -> 500MB data
- Case 3: 6-to-1 incast based on Case 1

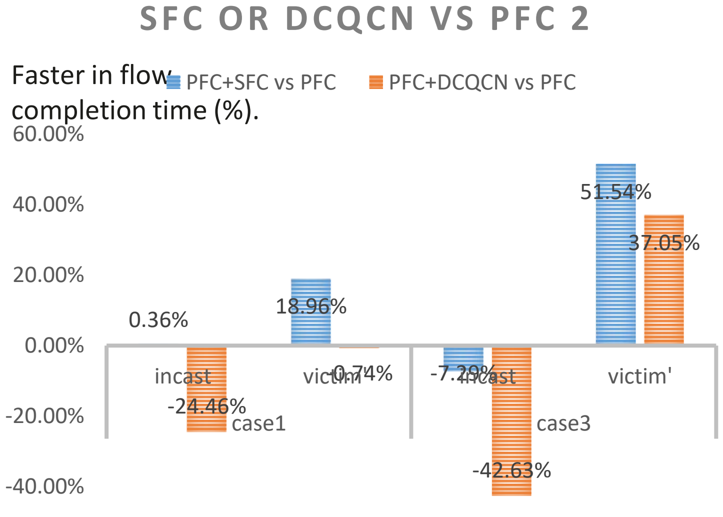


8 GPU per server, 128 GPU in total

Flow completion time (ms)		PFC only	PFC+SFC	PFC+DCQCN
case1	incast	238.5	237.6	444.9
	victim	355.6	288.1	690.3
case2	incast	2398.0	2355.5	3483.7
	victim	3738.1	2848.5	3431.8
case6	incast	24034.0	23476.0	27015.1
	victim	37527.8	28451.7	31901.6
case3	incast	418.9	449.5	585.9
	victim	533.2	258.4	304.625



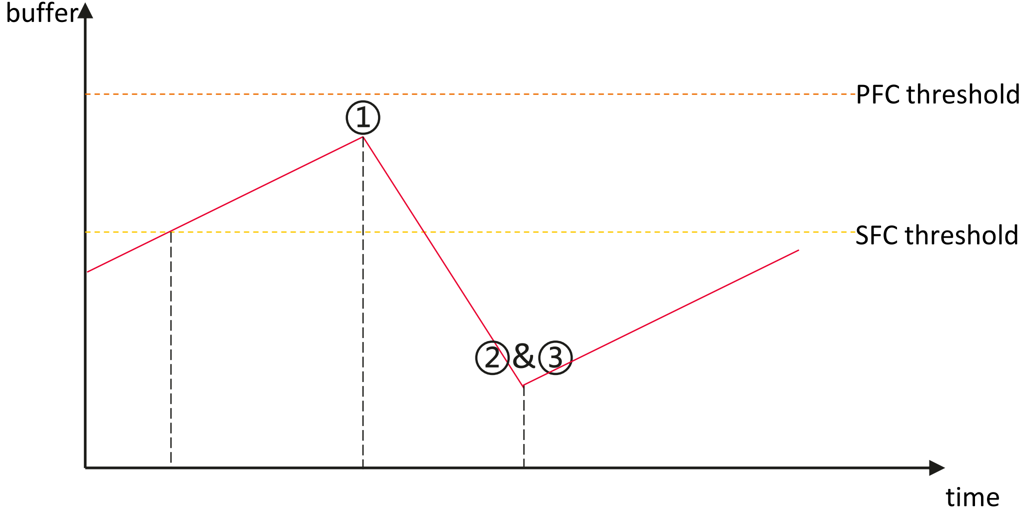
The longer the single flow duration, the better the performance of DCQCN.



More severe the incast, more improvement SFC and DCQCN provides the victim flow.

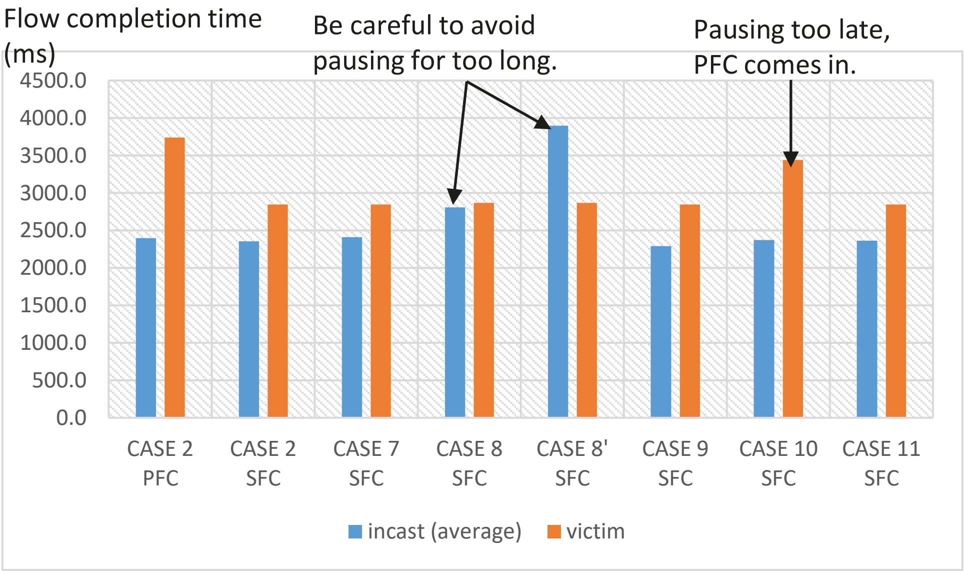
Simulation Part 2: SFC robustness

https://www.ieee802.org/1/files/public/docs2025/dw-chen-sfc-computation-simulation.pdf



- All settings based on case 2, where
  - > Buffer size: 400KB, PFC threshold: 380KB, SFC threshold: 200KB
  - > SFC pause time: 6us, should be within (4.8, 12.8)
- Case 7: pause longer (12us), but still within the right range.
- Case 8: pause too long (30us), SFC over-reacts.
  - > Case 8’: pause even longer (50us).
- Case 9: pause too short (3us), consuming more resources.
- Case 10: SFC threshold too high (260KB), PFC may come in. Similar to the result when SFC reacts too slowly.
- Case 11: SFC threshold too low (100KB), aggressive flow control.

	CASE 2	CASE 2	CASE 7	CASE 8	CASE 8'	CASE 9	CASE 10	CASE 11
incast	2398.0	2355.5	2410.9	2810.0	3898.5	2290.4	2355.5	2363.0
victim	3738.1	2848.5	2848.5	2868.9	2868.9	2848.5	2848.5	2848.5



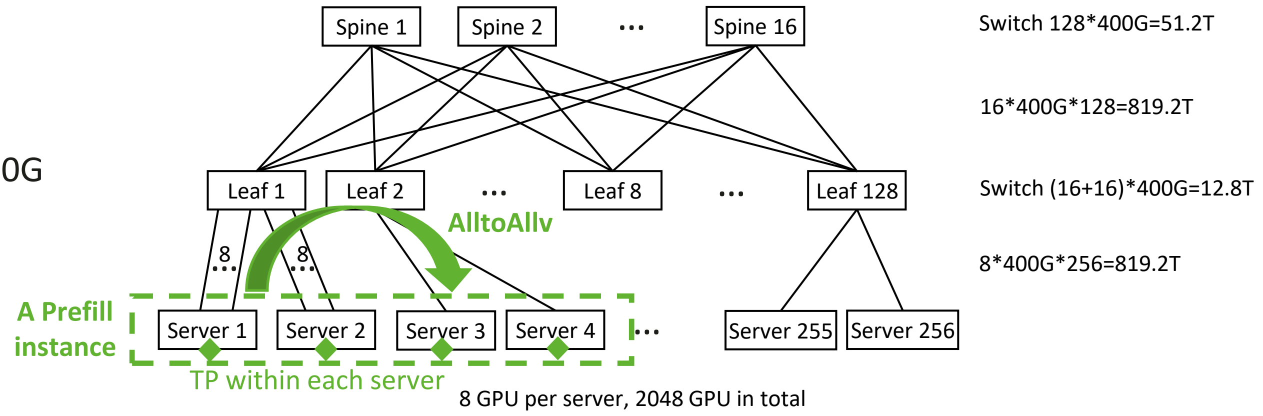
Be careful to avoid pausing for too long.

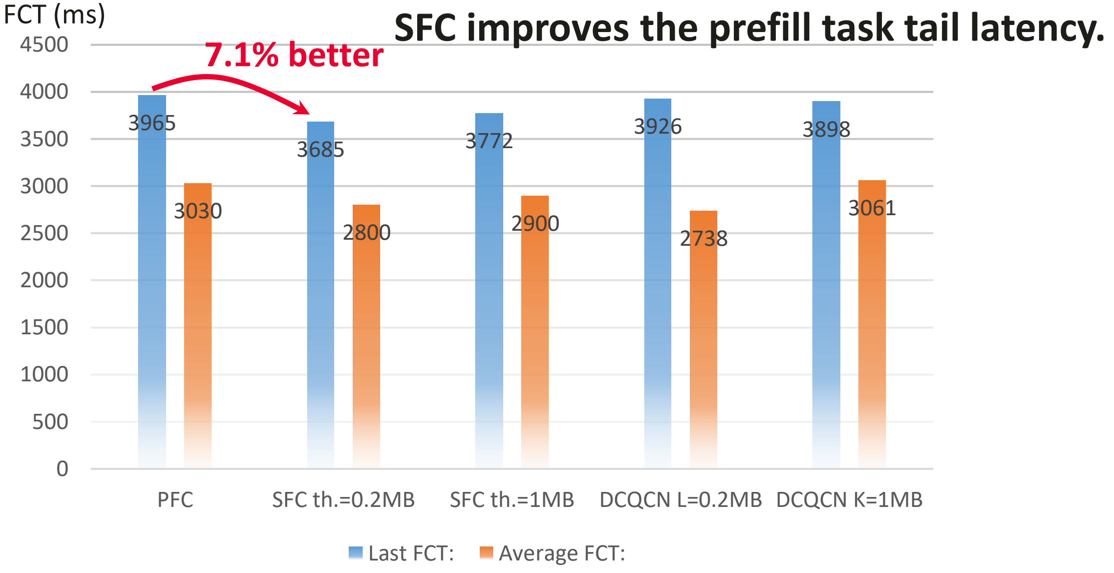
Pausing too late, PFC comes in.

The key is to avoid pausing for too long, as bandwidth under-utilization is unacceptable. All other situations are fine.

Simulation Part 3: DC backend network for AI inference

- Network 2-layer CLOS
  - > 16 spine switches, 128*400G. 128 Leaf switches, 32*400G
  - > 2048 GPU, 400G
- Basic settings
  - > Link delay: 150ns. Switch process (fixed) delay: 300ns
  - > Max. packet size: 4KB
  - > Buffer size: 2MB. PFC threshold: 1.9MB.
- Flow settings (Prefill in inference) and explanation
  - > A prefill instance has 32 GPUs. 8 experts.
  - > Each AlltoAllv, all 32 GPUs simultaneously send data to 8 other GPUs.
  - > Data size: 14.7MByte (Relevant parameters: batch size, sequence length, hidden size, Tensor Parallelism, Data Parallelism).
  - > The next round of AlltoAllv will start after transmission of all these 32*8 flows have completed.
  - > The **last flow completion time (FCT) matters**!



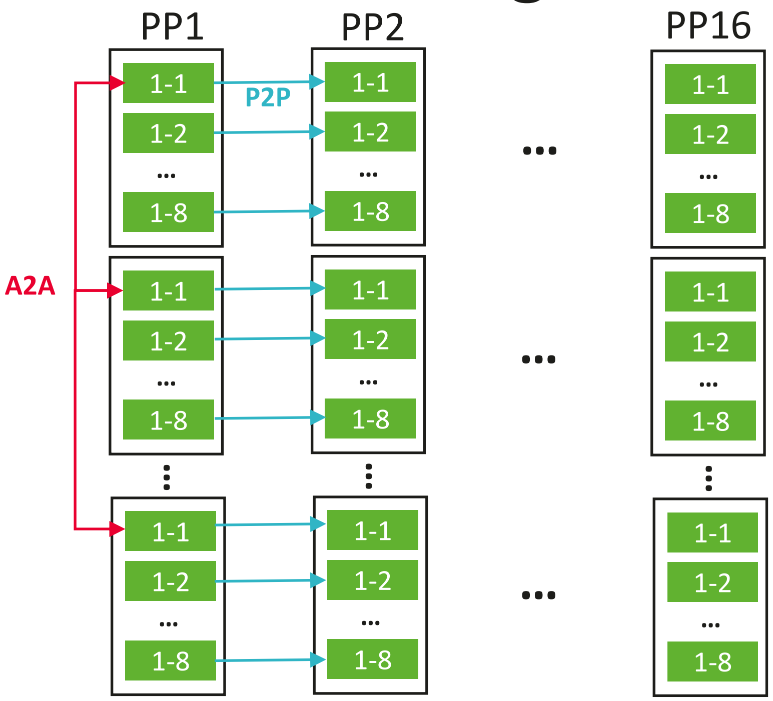


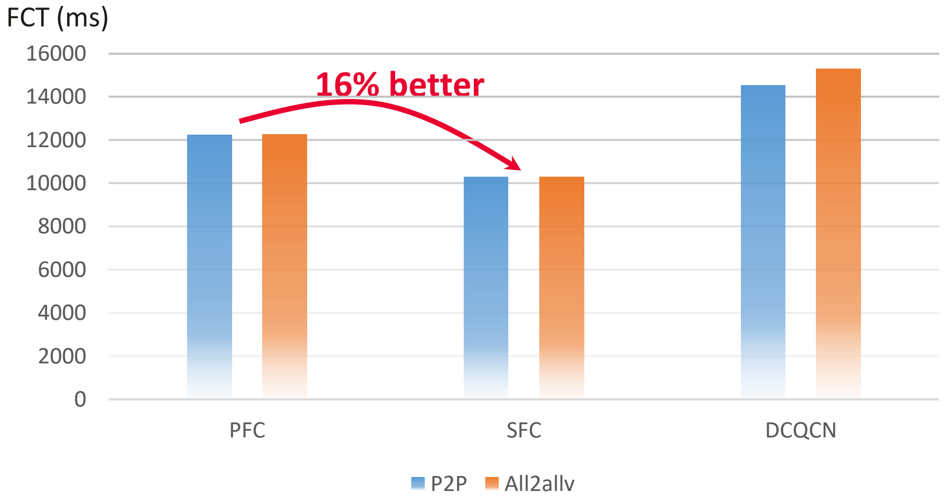
SFC improves the prefill task tail latency.

For a 2048-GPU cluster, Prefill : Decode = 1:1, communication : computation = 1:3, save equivalently 18 GPUS (improved by 0.89%). Global electricity generation to supply DCs is 460TWh in 2024. Saving 0.89% means 4TWh, ≈800M USD ($0.2/kWh).

Simulation Part 3’: DC backend network for AI training

- Simulator: Same as in Page 8 of https://www.ieee802.org/1/files/public/docs2025/dw-chen-sfc-computation-simulation.pdf
- Topology: same as in Page 7, 128 servers, 1028 GPUs
- Flow settings (Training) and explanation:
  - > 16 Pipeline Parallelism (PP), with 8 servers in each PP.
  - > AlltoAllv: Select 8 experts limited to 4 servers within that PP, and then communicate through the same-numbered GPUs between servers.
  - > P2P: Send data to the same-numbered GPU of the same-numbered server in the next PP.
  - > Data size: AlltoAllv 51.375MB, P2P 58.7MB.
  - > The PP1 computes a batch of tokens and communicates through AlltoAllv, then sends data (P2P) to the next PP2. PP2 continues computing while PP1 starts compute the next batch of tokens.
  - > The **last flow completion time (FCT) matters**!





Discuss and next step?

Draft status and Proposal for the next step

- The latest individual text and its brief can be found at https://www.ieee802.org/1/files/public/docs2025/dw-chen-individual-text-0325-v03.pdf and https://www.ieee802.org/1/files/public/docs2025/dw-chen-text-status-and-todos.pdf
- The introducing (clause 1-6) and the concept and component description (52.1-52.4) parts are almost there.
- The management part (12, 48) can be handled last.
- Clause 52.5 is the meat. For **next step**:
  - > Update and finalize a first but complete version of SFCP procedure (52.5.2).
  - > Update Encoding (52.5.3) based on what has been proposed in this contribution.
  - > Revise Variables (52.5.1) accordingly.
  - > Add Buffer requirements for SFC (Annex Y) based on the calculation given in https://www.ieee802.org/1/files/public/docs2025/dw-chen-sfc-computation-simulation.pdf
- Then, **take the draft to a Task Group ballot**. (Nov.2025?)