Welcome to
Hot Chips 32!

Organizing Committee

Alisa Scherer

General Chair

Conference Sponsor: IEEE Technical Committee on Microprocessors and Microcomputers

Conference Details & Stats

- Welcome to our 1st Virtual Conference!

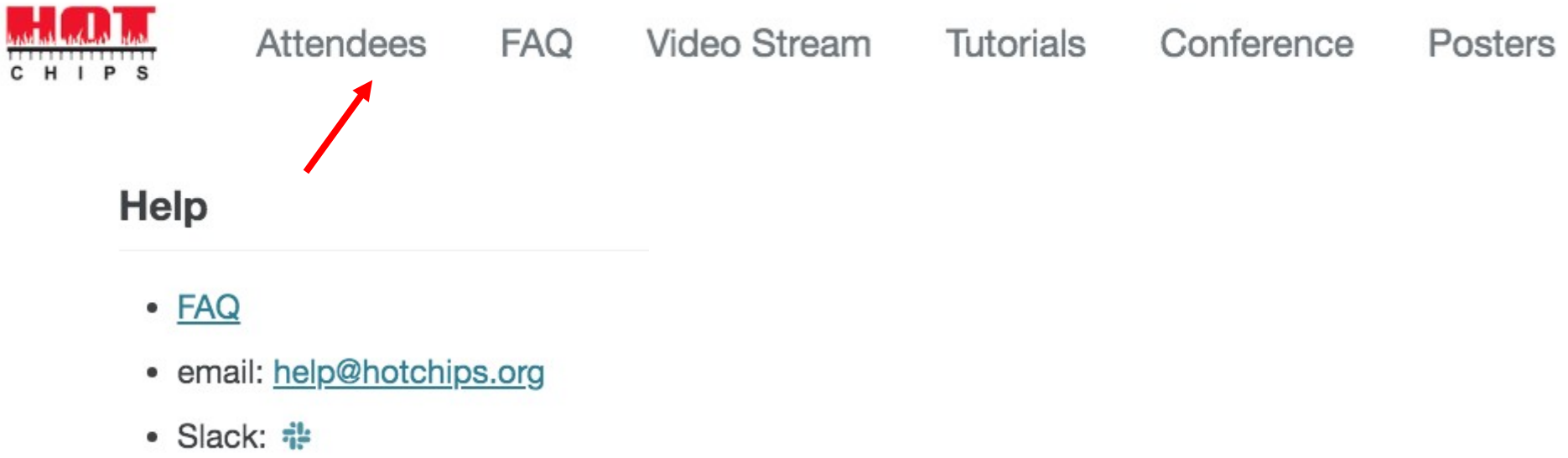
- Registration
  - Record numbers!!!

- Sponsorship
  - Please visit our sponsors' tables in the courtyard:
  - https://hc32.hotchips.org/sponsors_courtyard/

Virtual Conference Details: Help

Website: https://www.hc32.hotchips.org/

- Get help with logins and Slack access



Virtual Conference Details: Features

Attending the Virtual Conference

There are three main features of this year's all virtual conference:

1. Live video streaming links of tutorials and conference presentations (streaming video player is also embedded in program pages)
    - Recorded videos of the talks/posters are available where you see this icon ex: posters
2. Slack for Q&A with speakers, poster presenters, sponsors and other attendees with this icon
3. This website with program pages and PDFs with this icon Detailed Attendee Help

Live Video

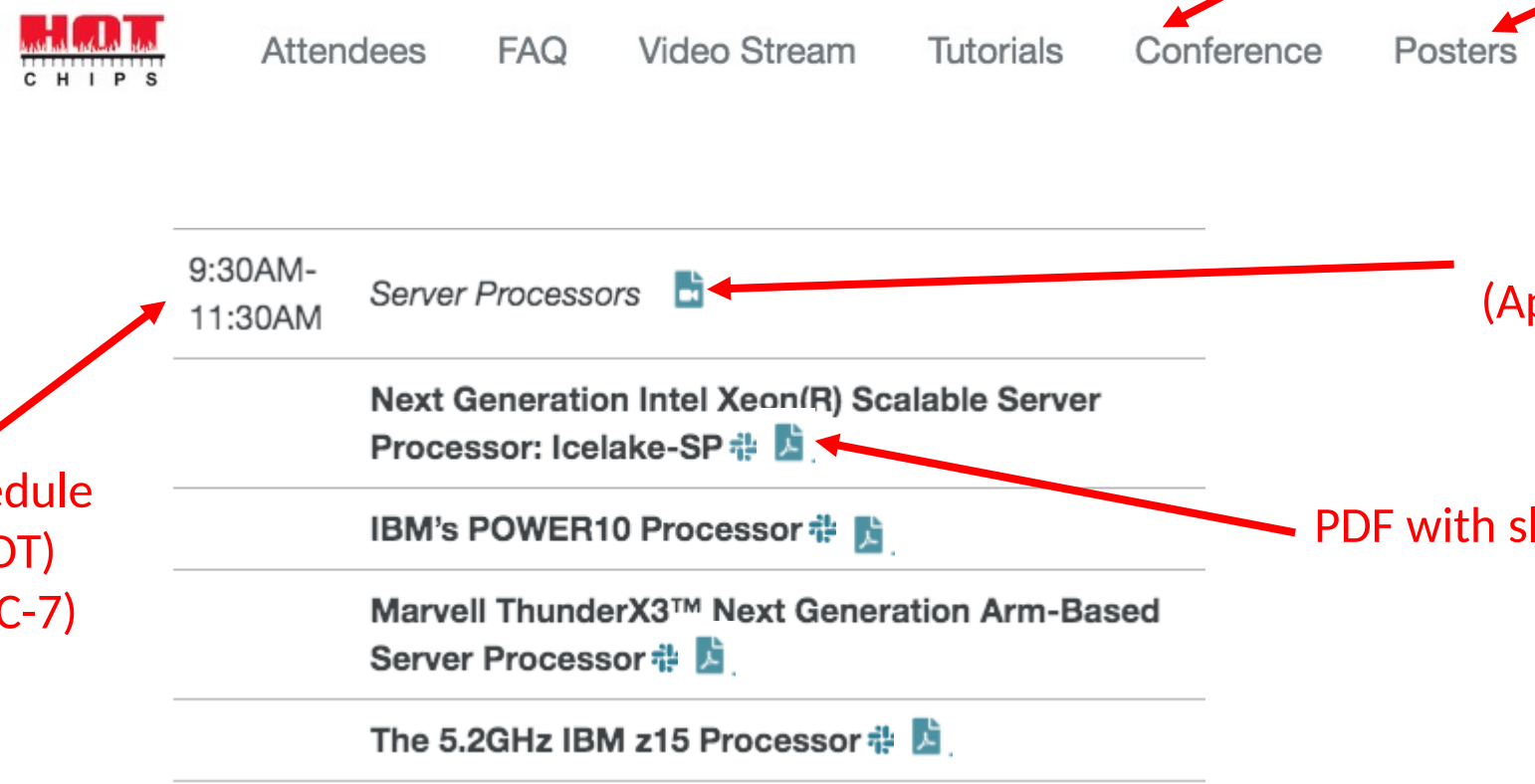
Recorded Video

Slack link

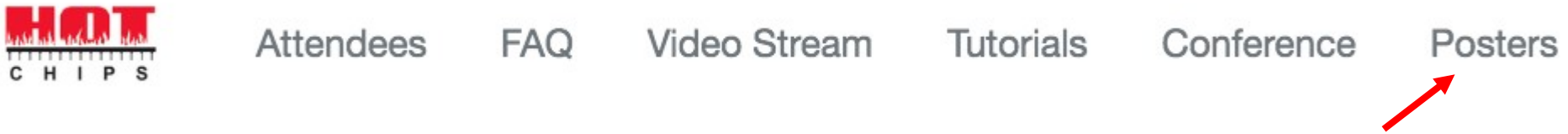
PDFs

Virtual Conference Details: Slides

- Use your personal password from the email to access Tutorials and Conference pages.



Virtual Conference Details: Posters



Hot Chips 32 selected 11 outstanding posters

- Visit them on the Posters tab
- PDFs for viewing posters
- Slack Channels for interaction with presenters
- Video links

Behind the scenes

- Volunteer-run conference
  - Steering Committee
  - Organizing Committee
  - Program Committee
  - Volunteers

Welcome to the
Hot Chips 32 Program

Priyanka Raina (Stanford)
Cliff Young (Google)

Hot Chips 32 Program Committee

Krste Asanovic, UC Berkeley/SiFive
Forest Baskett, NEA
Ian Bratt, ARM
Ron Diamant, Amazon
Pradeep Dubey, Intel
John Hennessy, Stanford/Google
Yoshio Masubuchi, Toshiba
Rob Ober, NVIDIA
Priyanka Raina, Stanford (co-chair)
John Sell, Microsoft
Sophia Shao, UC Berkeley
Alan Smith, UC Berkeley
Fred Weber
Ralph Wittig, Xilinx
Yuan Xie, UCSB/Alibaba
Misha Smelyanskiy, Facebook
Cliff Young, Google (co-chair)

Tutorials

- Machine Learning Training Scale Out
  - Scale Out Systems: NVIDIA, Google, and Cerebras
  - Scale Out Training Experiences: NVIDIA, Baidu, and Google

- Quantum Computing: UCSB, Google, IBM, Intel, and Microsoft

Keynotes

No Transistor Left Behind

Raja Koduri

Senior Vice President, Chief Architect, GM of Architecture, Graphics, and Software

Intel

Monday 2.00pm (PDT)

AI Research at Scale - Opportunities on the Road Ahead

Dan Belov

Distinguished Engineer

DeepMind

Tuesday 1.30pm (PDT)

Program Statistics and Sessions

88 abstract submissions
25 accepted talks

Monday

Server Processors

Mobile Processors

Edge Computing and Sensing

GPUs and Gaming Architectures

Tuesday

FPGAs and Reconfigurable Architectures

Networking and Distributed Systems

ML Training

ML Inference

Server Processors

9:30 AM - 11:30 AM	*Server Processors* Chair: Pradeep Dubey	
	Next Generation Intel Xeon(R) Scalable Server Processor: Icelake-SP	Irma Esmer Papazian, Intel
	IBM's POWER10 Processor	William Starke and Brian W Thompto, IBM
	Marvell ThunderX3™ Next Generation Arm-Based Server Processor	Rabin Sugumar, Marvell
	The 5.2GHz IBM z15 Processor	Anthony Saporito

Mobile Processors

12:00 PM - 1:00 PM	*Mobile Processors* Chair: Fred Weber	
	AMD Next Generation 7nm RyzenTM 4000 APU	Sonu Arora, AMD
	Inside Tiger Lake: Intel's Next Generation Mobile Client CPU	Xavier Vera, Intel

Edge Computing and Sensing

3:00 PM - 5:00 PM	*Edge Computing and Sensing* Chair: Krste Asanovic	
	Xuantie-910: Innovating Cloud and Edge Computing by RISC-V	Yu Pu, Alibaba
	A technical overview of the Arm Cortex-M55 and Ethos-U55: ARM's most capable processors for endpoint AI	Allan Skillman and Tomas Edso, ARM
	PGMA: A Scalable Bayesian Inference Accelerator for Unsupervised Learning	Glenn G. Ko, Harvard University

GPUs and Gaming Architectures

5:00 PM - 6:30 PM	*GPUs and Gaming Architectures* Chair: John Sell	
	NVIDIA's A100 GPU: Performance and Innovation for GPU Computing	Jack Choquette and Wishwesh Gandhi, NVIDIA
	The Xe GPU Architecture	David Blythe, Intel
	Xbox Series X System Architecture	Jeff Andrews and Mark Grossman, Microsoft

FPGAs & Reconfigurable Architectures

8:30 AM - 10:00 AM	*FPGAs and Reconfigurable Architectures* Chair: Ralph Wittig	
	Agilex Generation of Intel FPGAs	Ilya Ganusov and Mahesh Iyer, Intel
	Xilinx Versal Premium Series	Martin Voogel, Yohan Frans and Matt Ouellette, Xilinx
	Compute Substrate for Software 2.0	Ljubisa Bajic and Jasmina Vasilijevic, Tenstorrent

Networking and Distributed Systems

10:30 AM - 12:30 PM	*Networking and Distributed Systems* Chair: Yuan Xie	
	Tofino2 – A 12.9Tbps Programmable Ethernet Switch	Anurag Agrawal and Changhoon Kim, Intel/Barefoot
	Pensando Distributed Services Architecture	Francis Matus, Pensando
	The DPU: A New Category of Microprocessor	Pradeep Sindhu, Fungible
	High-density Multi-tenant Bare-metal Cloud with Memory Expansion SoC and Power Management	Justin Song and Xiantao Zhang, Alibaba

ML Training

2:30 PM - 4:00 PM	*ML Training* Chair: Ron Diamant	
	Google's Training Chips Revealed: TPUv2 and TPUv3	Thomas Norrie and Nishant Patil, Google
	The Second Generation Cerebras Wafer Scale Engine	Sean Lie, Cerebras
	Manticore: A 4096-core RISC-V Chiplet Architecture for Ultra-efficient Floating-point Computing	Florian Zaruba, ETH Zurich

ML Inference

4:30 PM - 6:30 PM	*ML Inference* Chair: Forest Baskett	
	Baidu Kunlun – An AI Processor for Diversified Workloads	Jian Ouyang, Baidu
	Hanguang 800 NPU – The Ultimate AI Inference Solution for Data Centers	Yang Jiao, Alibaba
	Silicon Photonics for Artificial Intelligence Acceleration	Carl Ramey, Lightmatter

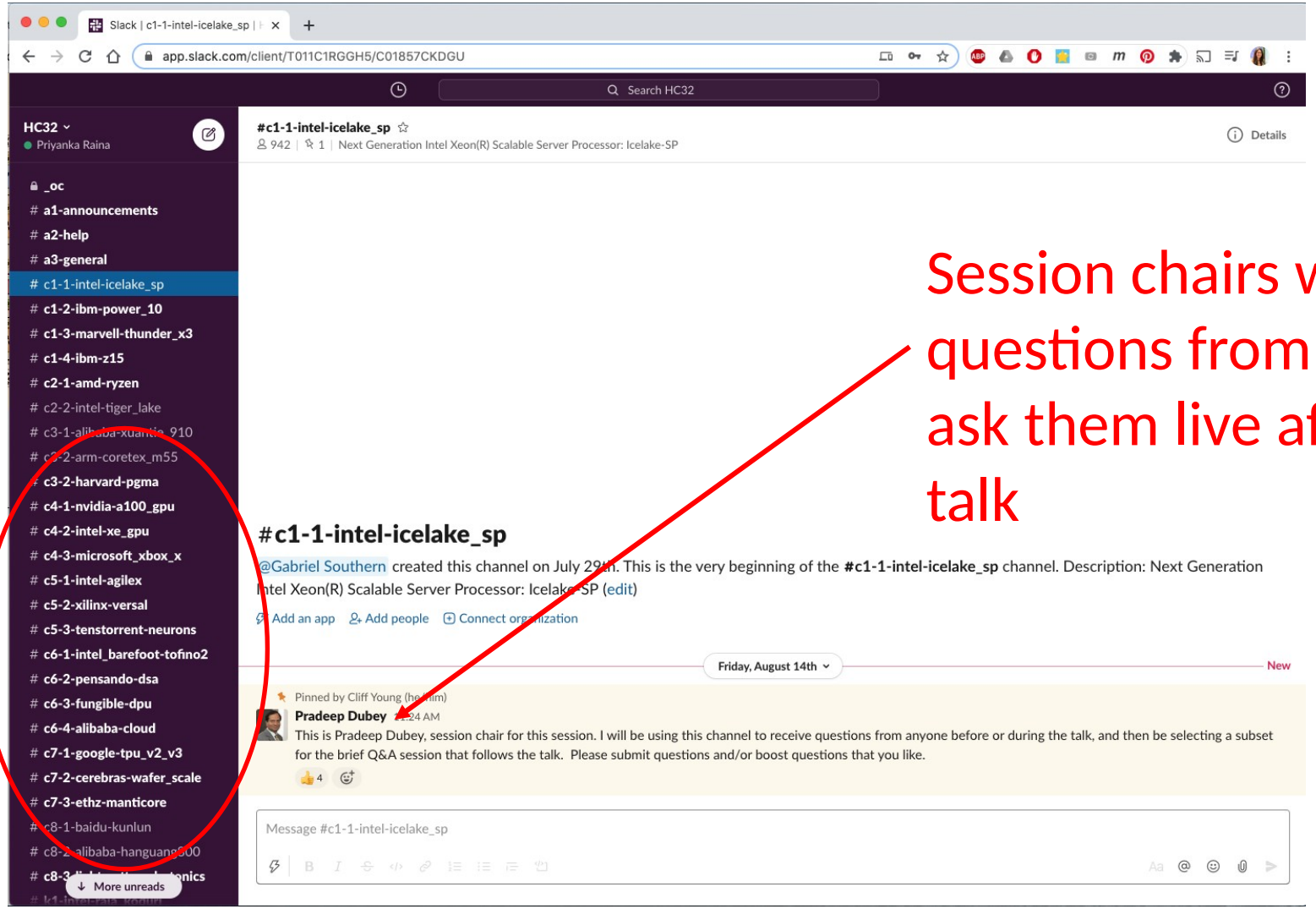
Q&A Process

Q&A for all sessions will happen through the HotChips Slack workspace: hc32.slack.com

You should have received the link to join Slack in an email

Q&A Process Through Slack

Each talk has a dedicated slack channel



Session chairs will pick questions from Slack and ask them live after each talk

Enjoy Hot Chips!