Redefining Voice Processing with Habana Gaudi

Exploring the Next Generation of Voice Processing Techniques

Rita Singh

Center for Voice Intelligence and Security
Carnegie Mellon University http://cvis.cs.cmu.edu

Presented at the
Intel oneAPI Dev Summit for AI and HPC
December 5, 2023

Topics

1. Habana Gaudi-2 hardware highlights
2. Introduction to voice processing model architecture
3. Alignment with hardware capabilities

Hardware Architecture Overview

Key Components

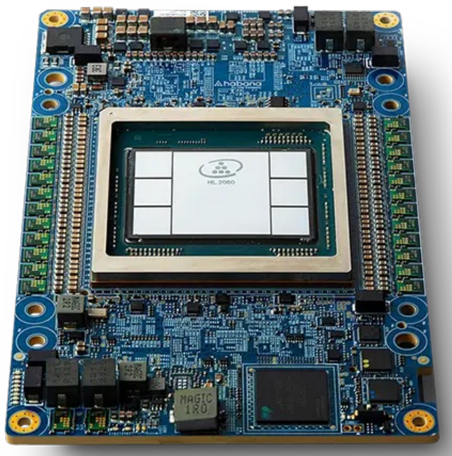
- Intel Xeon CPUs
- Habana Gaudi2 HPUs
- DDR5 RAM
- PCIe Gen5
- NVME SSD
- QSFP-DD

https://habana.ai/products/networking/

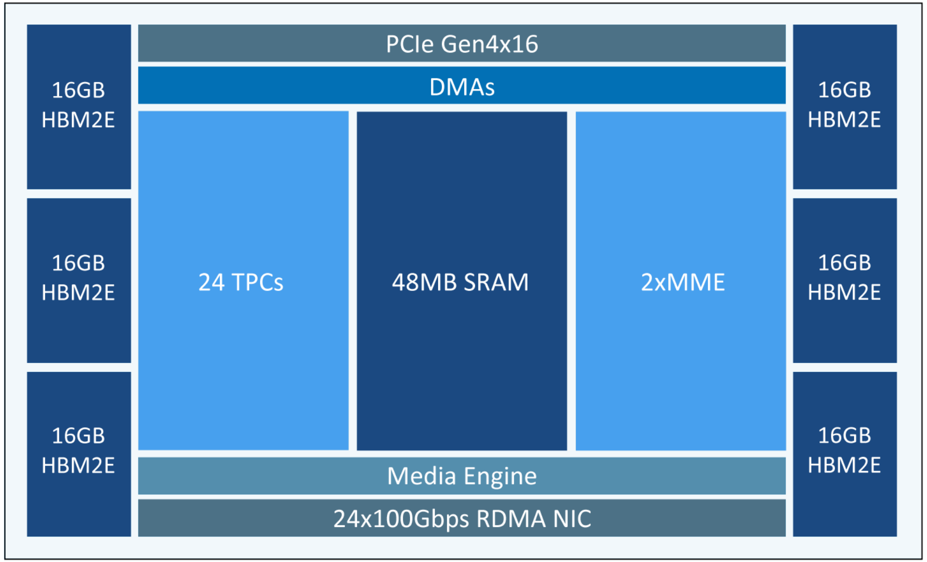
CPU Specifications - Intel Xeon CPU Max 9480

- 56 cores, 112 threads, Base 1.90 GHz, Turbo 3.50 GHz.
- 4 UPI links at 16 GT/s.
- Supports DDR5 4800 MT/s, 64 GB HBM.

Habana Gaudi2 HPU Information

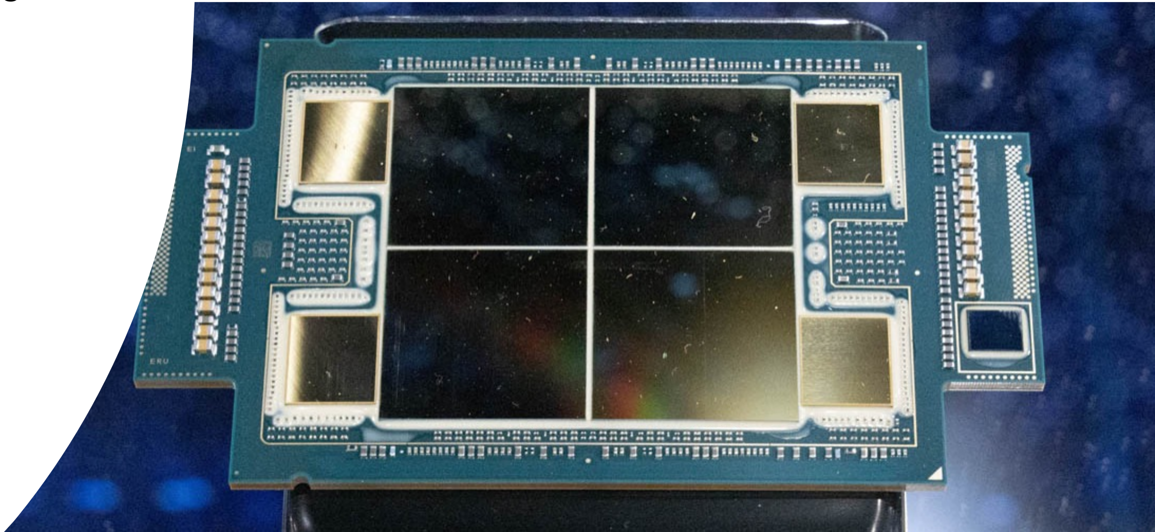


- 24 Tensor Processor Cores.
- 96 GB HBM2E memory onboard.
- Dual matrix multiplication engines.



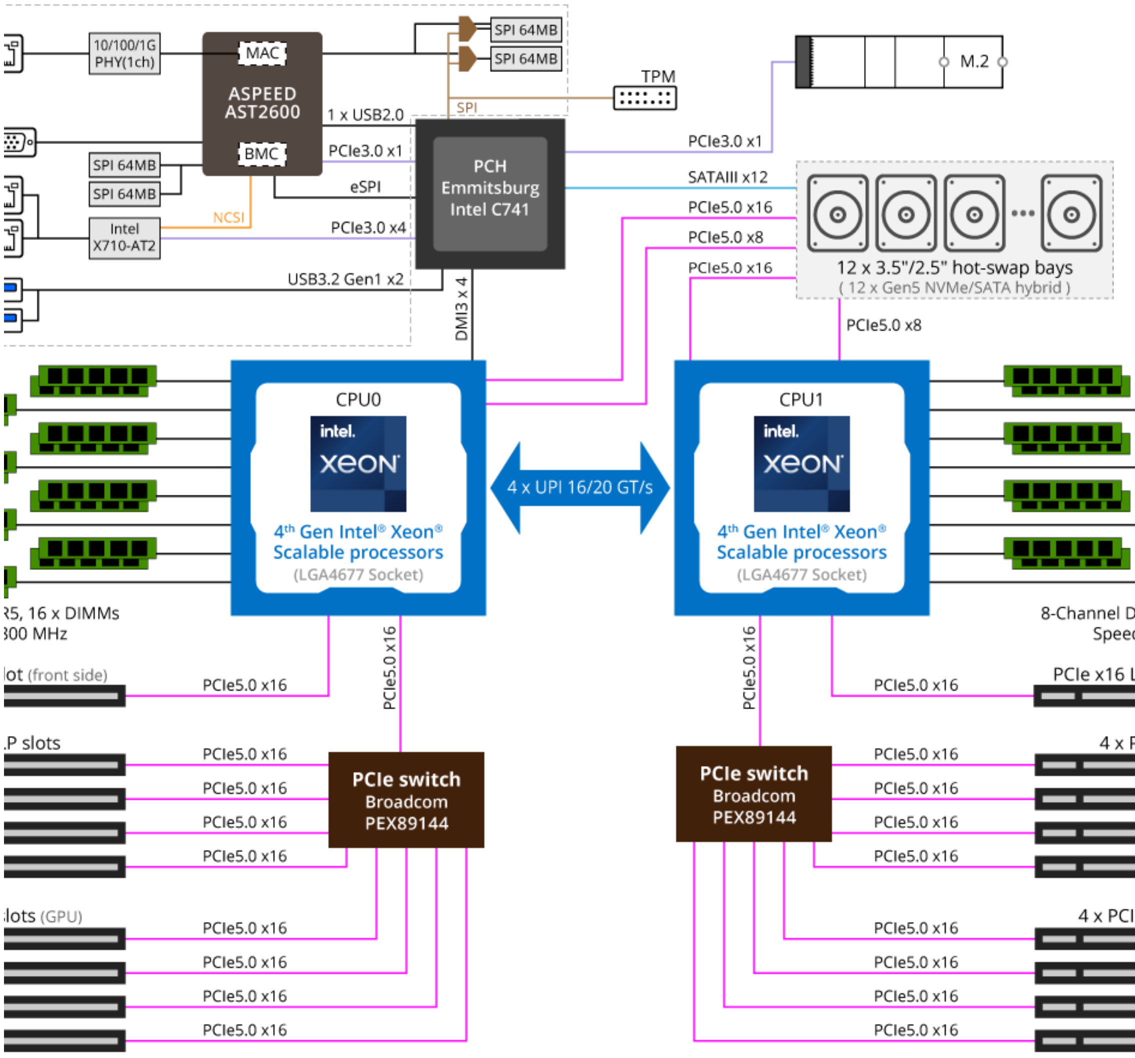
Memory and Storage Integration

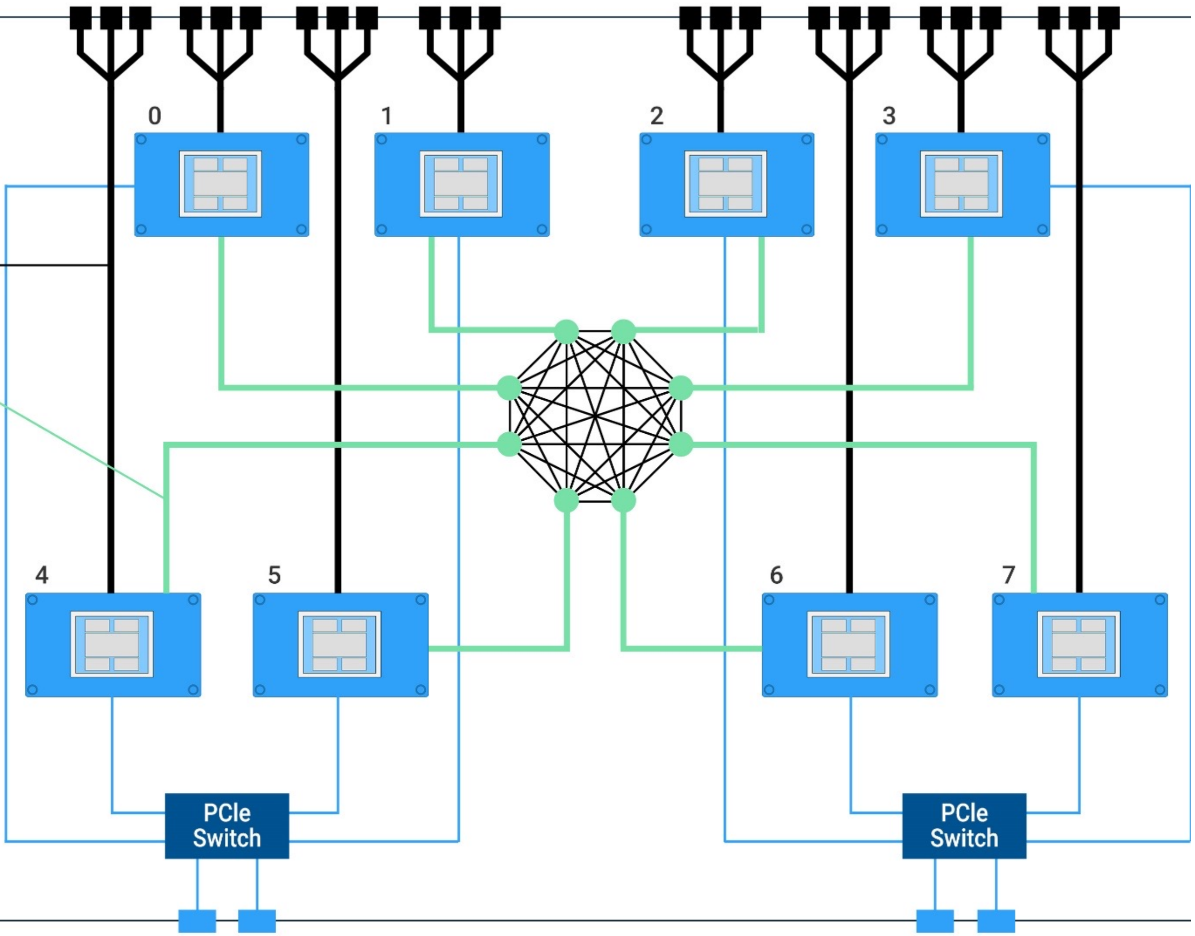
- RAM: DDR5, 8 channels, 4800 MT/s.
- HBM2E memory in HPU: 2.45 TB/s bandwidth.



PCIe and UPI Interconnects

- PCIe Gen5: 32 GT/s per lane, used for CPU-HPU, CPU-NIC connections.
- Intel UPI: 16 GT/s, connecting CPUs.





Networking: NIC and RDMA

- NIC Specs: 100 Gbps Ethernet.
- RDMA for inter-node communication.

Hardware Performance Analysis

- Bandwidth and latency analysis of each component.
- Potential bottlenecks in data flow.

Object1	Object2	Connection	Count	Latency	Bandwidth Per Connection	Total Bandwidth	Total Bandwidth in GB/s
CPU	CPU	Intel UPI	4	Few nanoseconds	16 GT/s	64 GT/s	8 GB/s
CPU	RAM	DDR5	8	~50-70 ns	4800 MT/s	38400 MT/s	48 GB/s
CPU	HBM	HBM2E	8	~1-2 ns	2.45 TB/s	19.6 TB/s	2450 GB/s
HPU	HBM	HBM2E	8	~1-2 ns	2.45 TB/s	19.6 TB/s	2450 GB/s
CPU	HPU	PCIe Gen5	8	Few microseconds	32 GT/s	256 GT/s	32 GB/s
CPU	NIC	PCIe Gen5	2	Few microseconds	32 GT/s	64 GT/s	8 GB/s
HPU	NIC	PCIe Gen5	8	Few microseconds	32 GT/s	256 GT/s	32 GB/s
HPU	HPU	PCIe Gen4	8	Few microseconds	16 GT/s	128 GT/s	16 GB/s

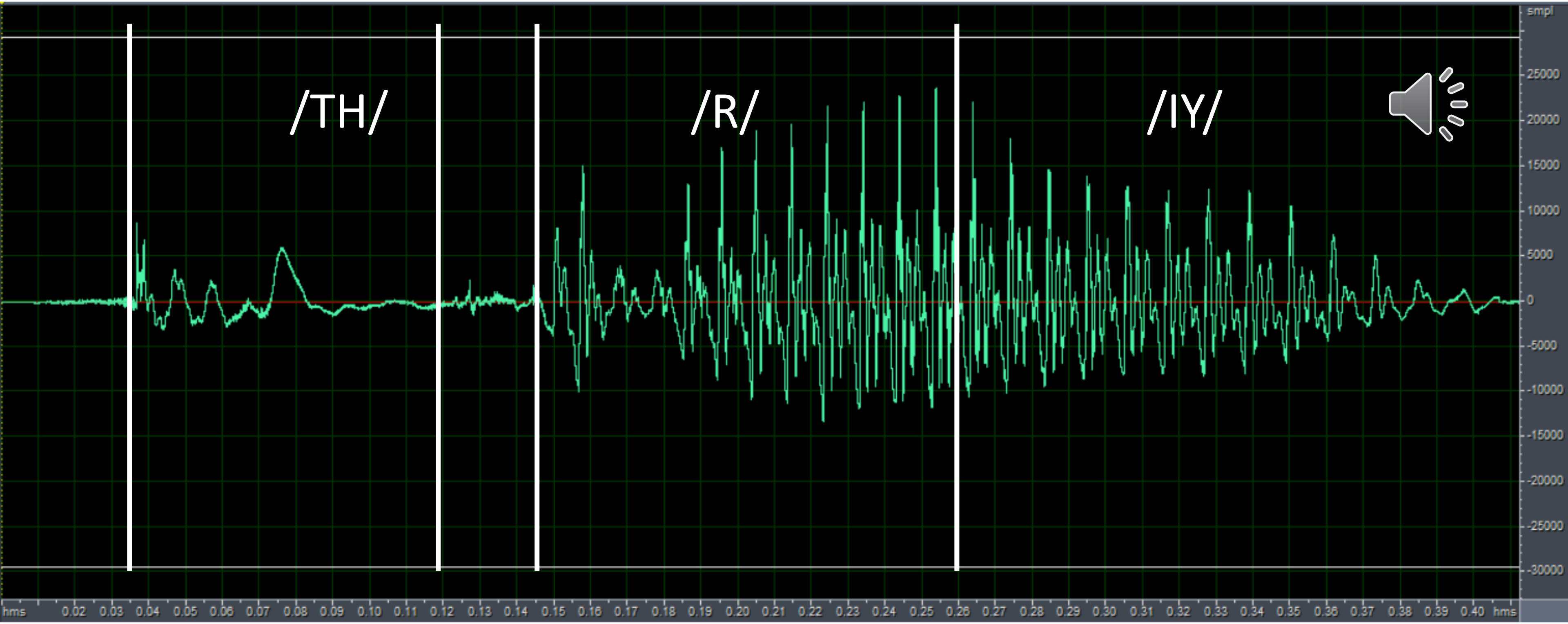
Topics

1. Habana Gaudi-2 hardware highlights
2. Introduction to voice processing model architecture
3. Alignment with hardware capabilities

Acoustic level: 100-250 glottal pulses/second

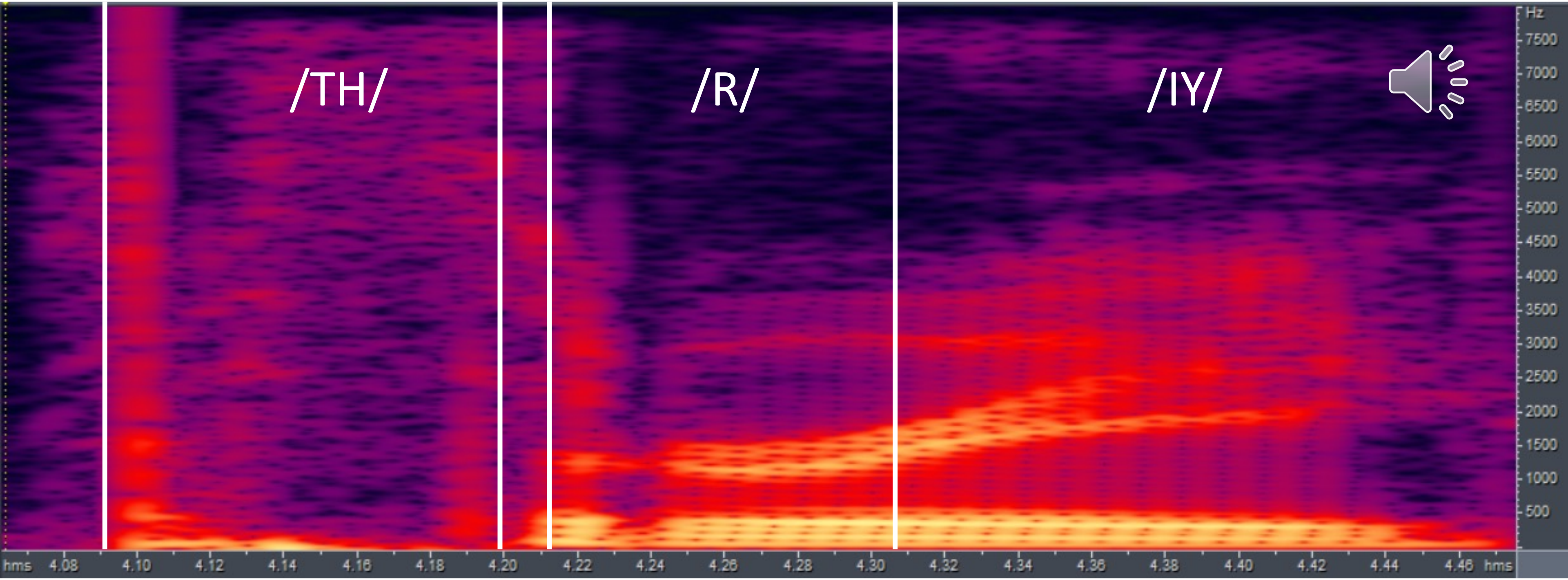
- Normally a 2.5 ms window would suffice (400 readings per second)
- The derivatives of the glottal flow waveform are important, leading to the requirement of 400-1000 readings per second

Sub-word level



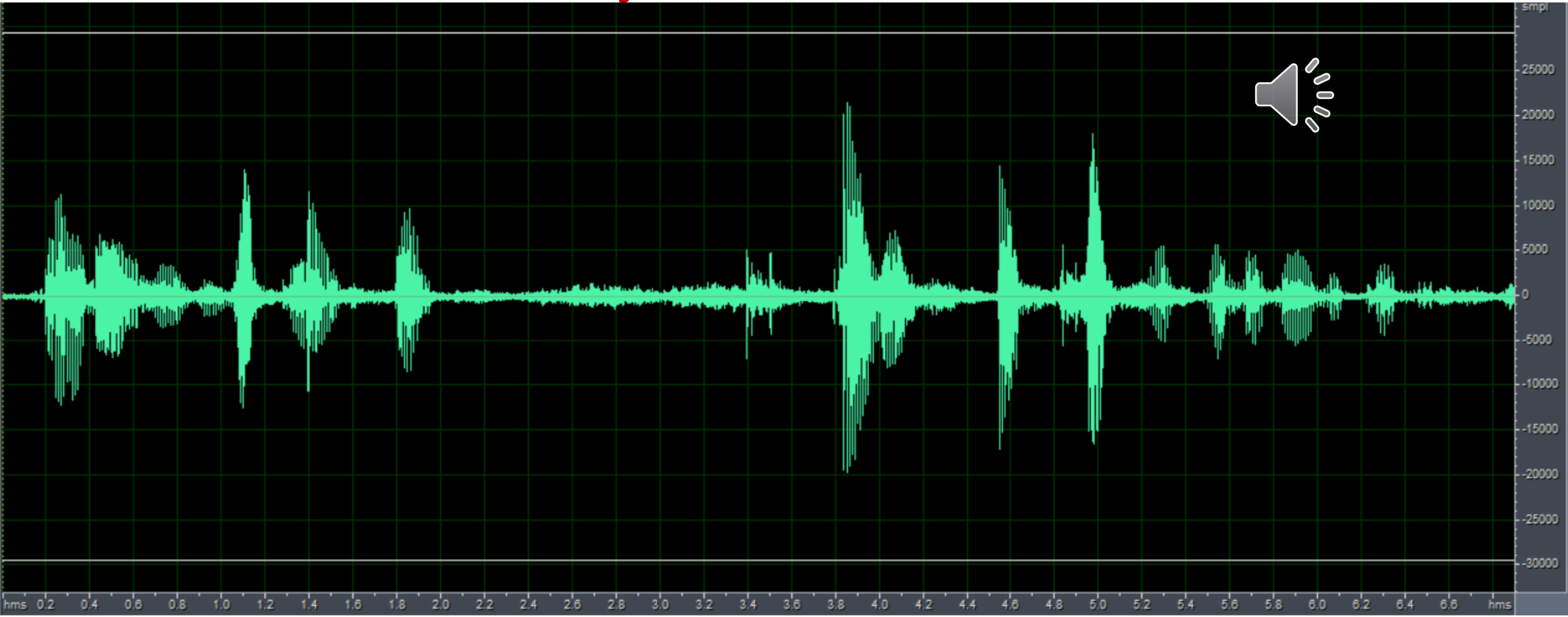
An example: Information in 0.412 seconds of speech

Sub-word level



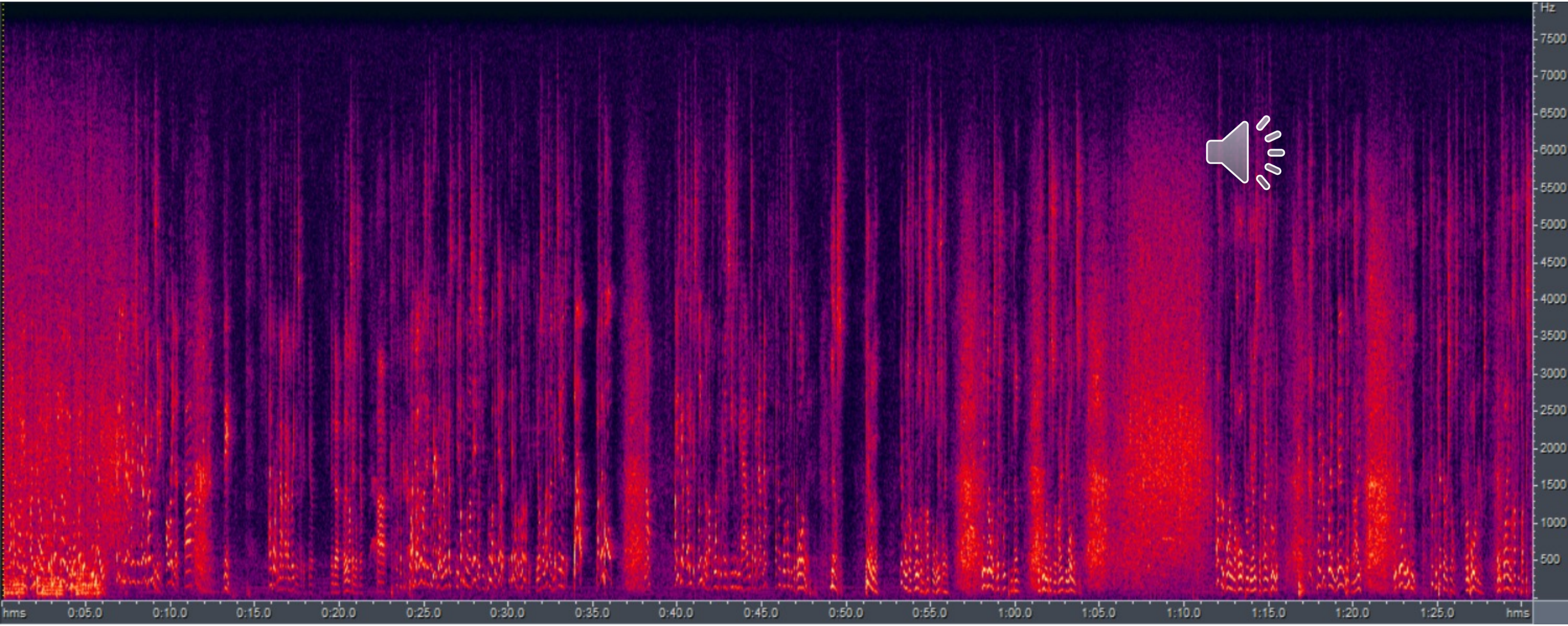
- An example: Information in 412 milliseconds of speech (needed to learn individual sounds with contextual effects for any language)
  - At least 15 seconds to gauge differences in prosody, identify dialect, language etc.

Syntactic level



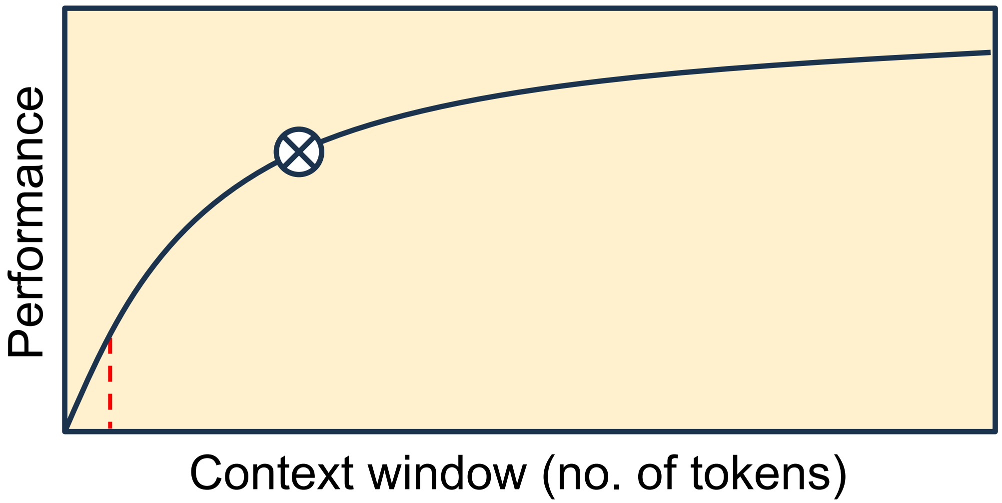
- An example: Information in 5.958 seconds of speech (needed to learn syntax of sentences, disfluencies, pause patterns etc for any language)
  - At least 30 seconds to gauge the full extent of syntactic patterns.

Semantic level



- An example: Information in 1min 30secs seconds of speech (needed to learn semantic interpretations from speech)
  - At least 2 mins to learn to identify subject, extend narrative, summarize etc.

Point of diminishing returns in LMs



Introduction to Voice Processing Challenges

- Granularities in speech processing

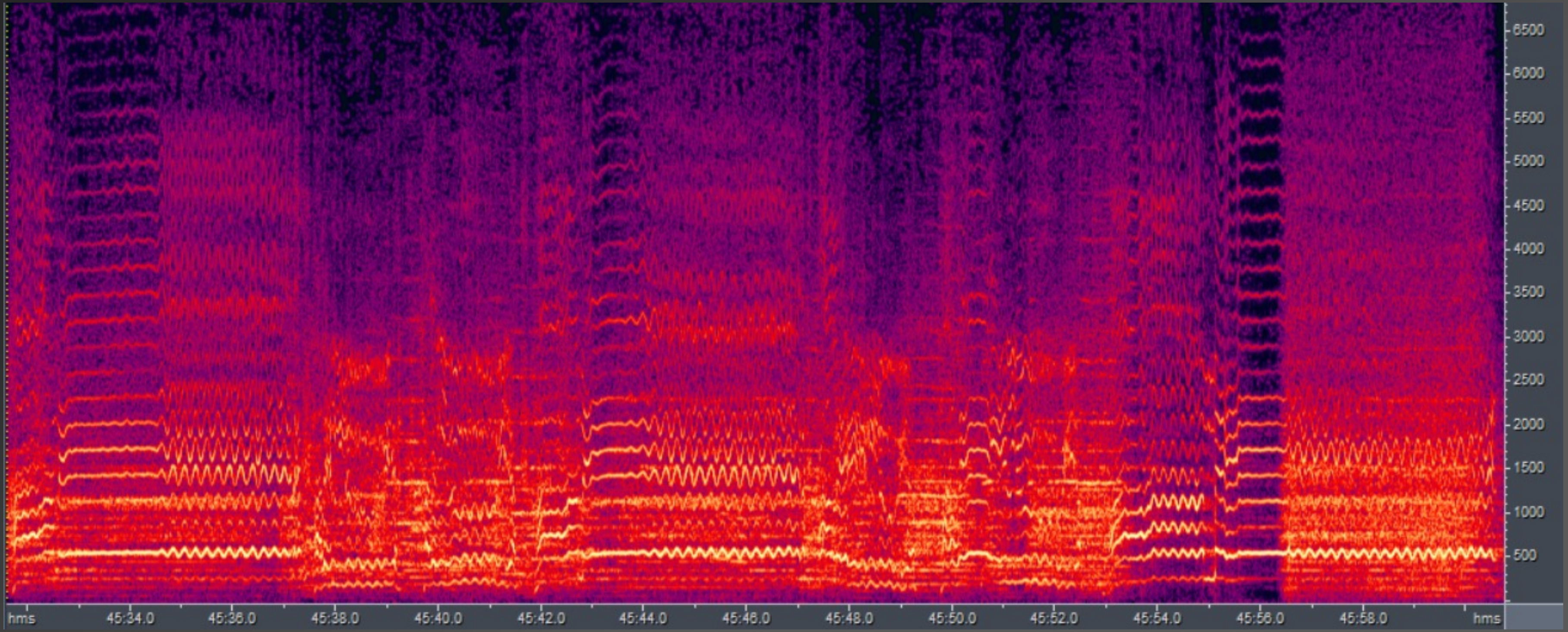
Acoustic 60 sec
400-1000 tokens per second
24000-60000 token context

Sub-word 300 sec
5-40 tokens per second
1500-12000 token context

Syntactic 900 sec
1.5-6 words per second
1350-5400 token context

Semantic 1800 sec
1.5-6 words per second
2700-10800 token context

Acoustic level again

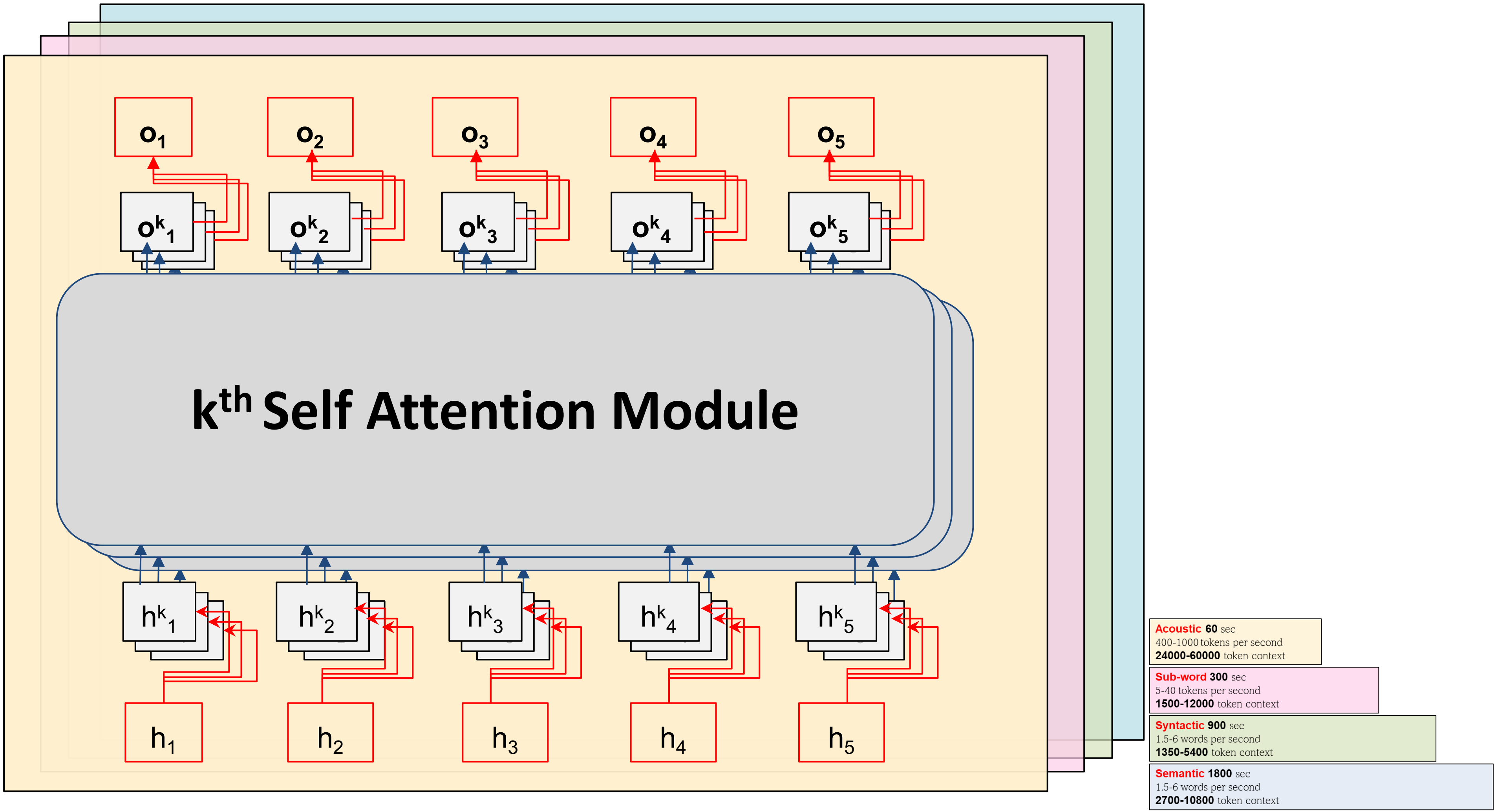


Cutting it close: Acoustic information in ~30 seconds of speech (in a song)

Acoustic level again

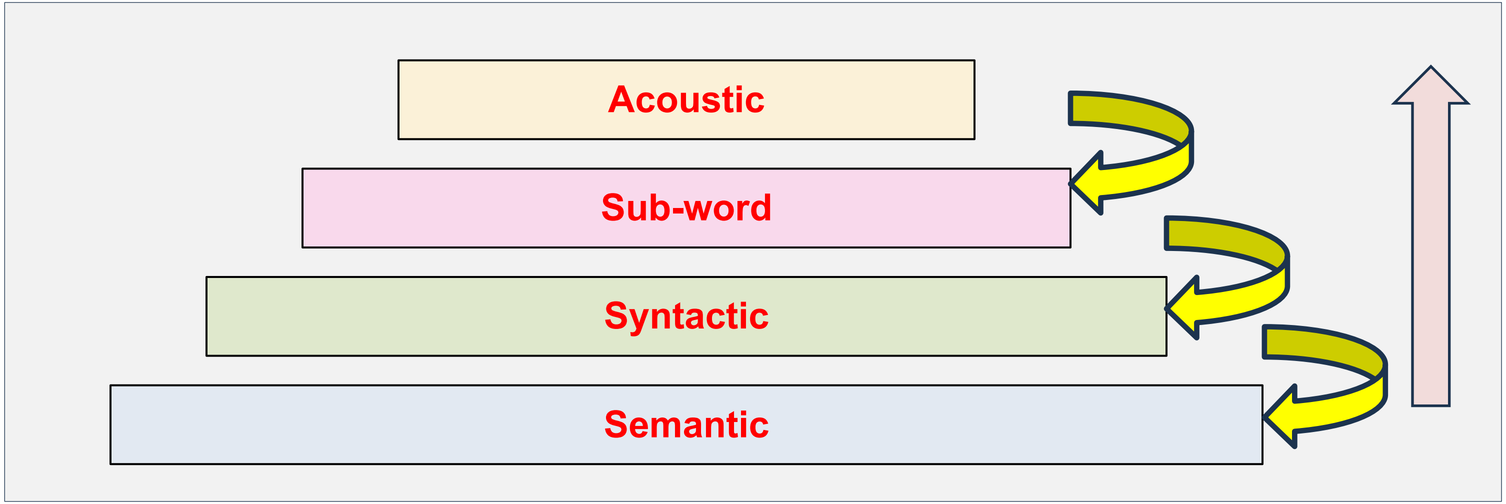
Nocturne, by Secret Garden

Purple People Eater, by Sheb Wooley (STEREO)



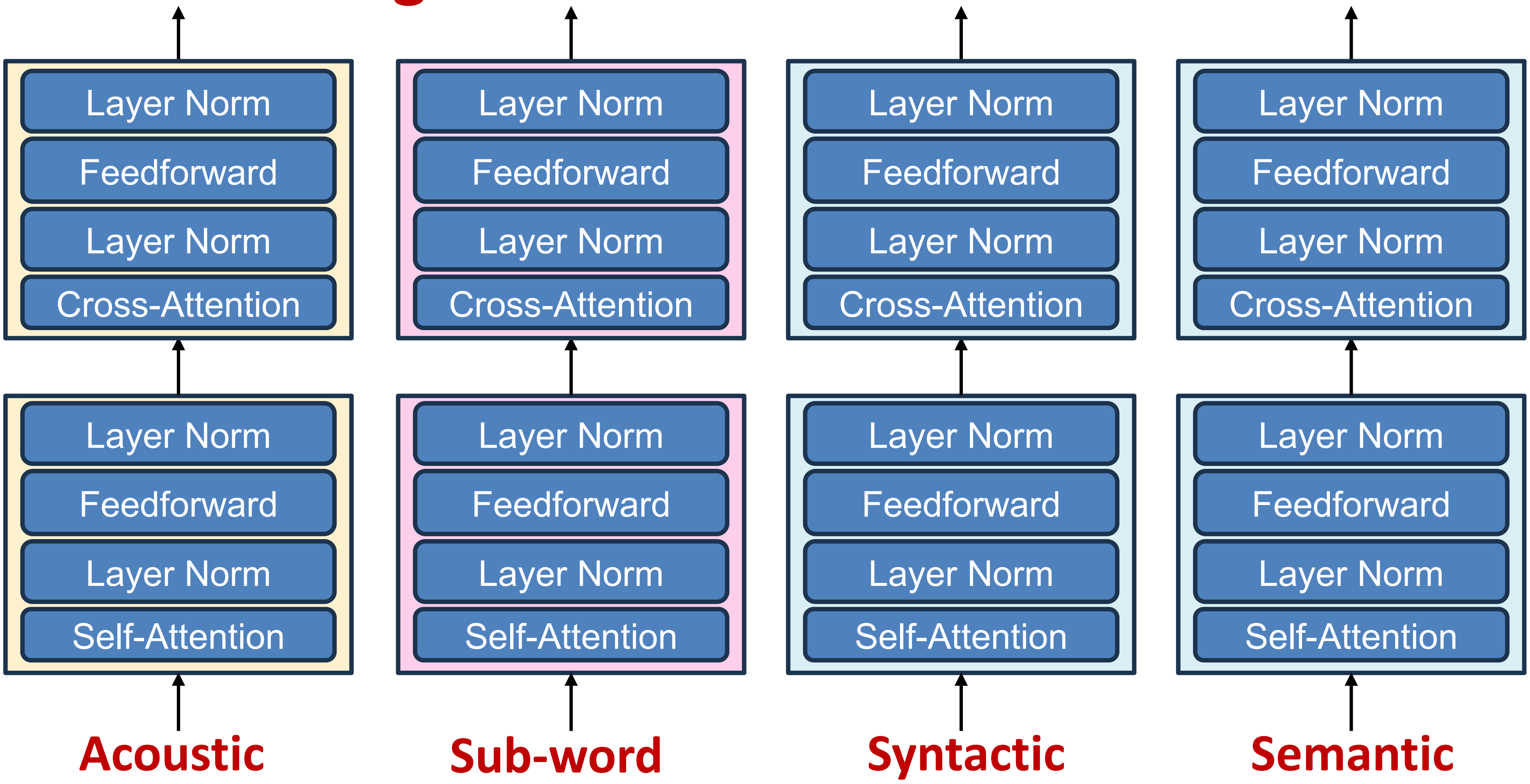
o$_1$
o$_2$
o$_3$
o$_4$
o$_5$
o$^k_1$
o$^k_2$
o$^k_3$
o$^k_4$
o$^k_5$
kth Self Attention Module
h$^k_1$
h$^k_2$
h$^k_3$
h$^k_4$
h$^k_5$
h$_1$
h$_2$
h$_3$
h$_4$
h$_5$
Acoustic 60 sec
400-1000 tokens per second
24000-60000 token context
Sub-word 300 sec
5-40 tokens per second
1500-12000 token context
Syntactic 900 sec
1.5-6 words per second
1350-5400 token context
Semantic 1800 sec
1.5-6 words per second
2700-10800 token context

Cross Attention



- Each layer "attends" to the lower one to update its representation
  - Cross attention "modules" are interspersed with self attention modules

High Level Model Architecture



Parameters for each self-attention block

- (8*3*[512 x 512/8])+(512*2) + ([512 x 2048] + [2048 x 512])+(512*2) = 2,885,632
  - 512 dimensional representation
  - 8 heads
  - Query, Key and Value sizes are 512/8
  - 2048 dimensional hidden representation for feedforward layer
  - 2 layer normalizations each with 512*2 parameters (512 dim scaling + 512 dim shift)
  - = 2,885,632
  - For 4 streams of information: 4* 2,885,632 = 11,542,528

Parameters for each cross-attention block

- (8*3*512*512/8)+(2*512*2048)+(512*4) = 2,885,632
  - 512 dimensional representation
  - 8 heads
  - Query, Key and Value sizes are 512/8
  - 2048 dimensional hidden representation for feedforward layer
  - 2 layer normalizations each with 512*2 parameters (512 dim scaling + 512 dim shift)
  - = 2,885,632
  - For 3 streams of information: 3* 2,885,632 = 8,656,896

Total for architecture with 50 alternate SA and CA blocks

- 50*(11,542,528+8,656,896) = 1,009,971,200
  - 512 is not enough for speech, need to use 1024 - 8192 dimensions
  - Total number of parameters can go up to 16,159,539,200 parameter

Computation in the full transformer (100 blocks)

1. Self-attention weights for one block: 10,800 * 10,800: *100 = 11,664,000,000 **+**
2. Cross-attention weights for one block = 0 **+**
3. 1 FLOP per parameter * 10,800 * 1,009,971,200 = 1.0907689e+13 FLOP
4. *This is for an 1800 second window*
5. *Min. req. for real-time processing* = **6066307222.22** FLOP/sec

Semantic 1800 sec
1.5-6 words per second
2700-10800 token context

1. Self attention weights for one block: 5400 x 5400 : *50 = 1,458,000,000 **+**
2. Cross-attention weights for 1 block = 5400*10800 : *50 = 2,916,000,000 **+**
3. 1 FLOP per parameter * 5400 * 1,009,971,200 = 5.4538445e+12 FLOP
4. *This is for a 900 second window*
5. *Min. req. for real-time processing* = **6064687222.22** FLOP/sec

Syntactic 900 sec
1.5-6 words per second
1350-5400 token context

1. Self attention weights computation: 12000 x 12,000 : *50 = 7,200,000,000 **+**
2. Cross-attention weights computation = 12000 * 5400 : *50 = 3,240,000,000 **+**
3. 1 FLOP per parameter * 12,000 * 1,009,971,200 = 1.2119654e+13 FLOP
4. *This is for a 300 second window*
5. *Min. req. for real-time processing* = **40433646666.7 FLOP**/sec

Sub-word 300 sec
5-40 tokens per second
1500-12000 token context

1. Self attention weights computation: [60000 x 60000] *50 = 180,000,000,000 **+**
2. Cross-attention weights computation= [60000 x 12000] *50 = 36,000,000.000 **+**
3. 1 FLOP per parameter * 60000 * 1,009,971,200 = 6.0598272e+13 FLOP
4. *This is for a 60 second window*
5. *Min. req. for real-time processing* = **1.0135712e+12**FLOP/sec

Acoustic 60 sec
400-1000 tokens per second
24000-60000 token context

1.0661358e+12 FLOPs

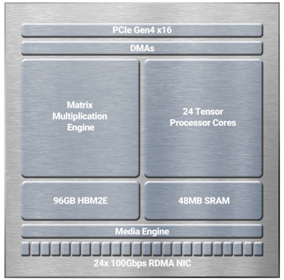
Topics

1. Habana Gaudi-2 hardware highlights

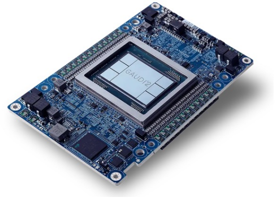
2. Introduction to voice processing model architecture

3. Alignment with hardware capabilities

Considerations: Parallelism & Accelerators



Gaudi2 Accelerator Architecture



- 3D Parallelism
  - Data, Pipeline and Tensor
- Support for Mixed Precision (float32, float16)
- Accelerating with Deepspeed
  - ZeRO (Optimizer Offloading, Gradient Offload, Model Parameter Partitioning)
- Gaudi-2 supports DDP (distributed data parallel) training using PyTorch
- Gaudi-2 supports TF.distribute training scheme in TensorFlow

Data, Tokens and Scaling

- 100k hours of speech with 400 tokens extracted for each second of speech amounts to a total of 144T tokens. This requires:
  - **For Audio:** In linear pcm mono 16kHz sampling rate format would need 10.48 TB of storage
  - **For Tokens:** 16 bit sample depth, 512 dimensional vectors: 134,108.16 TB
- In comparison, Llama2 uses only 3T tokens for training with text
  - Thus to build the next gen voice systems, we would process 48x more tokens for just 100k hours of data
- We have over 1M hours of speech data available today
  - more tokens and hours needed!
  - Over 1 Exabyte needed…

Considerations: Memory & Communications

Memory and Bandwidths control training efficiency with pipeline parallelism

- **Memory:**
  - Each HPU has 96 GB HBM VRAM
    - Each 8-HPU Node has 768 GB HBM VRAM
  - Total tensors for forward and backprop assuming 1000 fps and 4 byes per sample
    - 60000*60000 context * 512 dimensions * 4 bytes = 7,372,800,000,000 bytes = 6.7 TB
  - Total tensors for forward and backprop assuming 400 fps and 2 bytes per sample
    - 24000*24000 context * 512 dimensions * 2 bytes = 536 GB
    - The 8-HPU node is sufficient

Considerations: Memory & Communications

Memory and Bandwidths control training efficiency with pipeline parallelism

- **Bandwidth:**
  - Memory bandwidth of HPU processor controls transfer of data between CPU and HPU transfer of data between CPU and HPU
  - Bandwidth of connection between processors governs sharing of computations across HPU units for different parts of the model
  - Total tensors that need transferred for forward and backprop assuming 1000 fps
    - [(60000 acoustic * 12000 subword)*512 + (12000 subword *5400 syntactic)*512 + (5400 syntactic*10800 semantic)*512] * 4 bytes = 1.7267e+12 bytes = 1726 GB
  - Total tensors that need transferred for forward and backprop assuming partial 400 fps
    - [(24000 acoustic * 1500 subword)*512 + (1500 subword *5400 syntactic)*512 + (5400 syntactic*10800 semantic)*512]* 4 bytes = 2.0756e+11 bytes = 207.56 GB
  - 900 GB/s across HPUs

Hardware-Model Synergy for Optimized Training

- Each hardware component's capabilities align with the computational demands of the model.
- You can access Gaudi 2 through Intel® Developer Cloud.

Thank you!