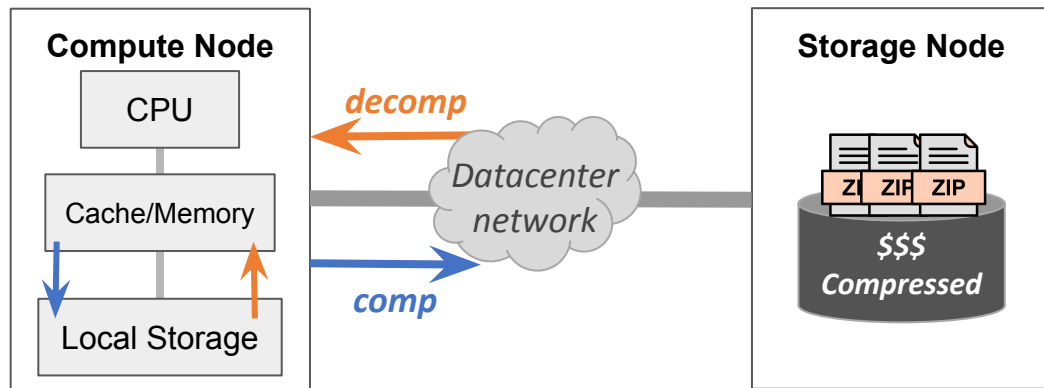


# The Free Lunch Has a Queue: Characterizing On-chip Compression Accelerators for Analytics

Yi Jiang, Antonio Boffa, Hamish Nicholson, Anastasia Ailamaki

# Compression is key to analytics performance



- + Reduce storage footprint
- + Minimize I/O and data movement overhead

Datacenter tax: 32-40% CPU cycles

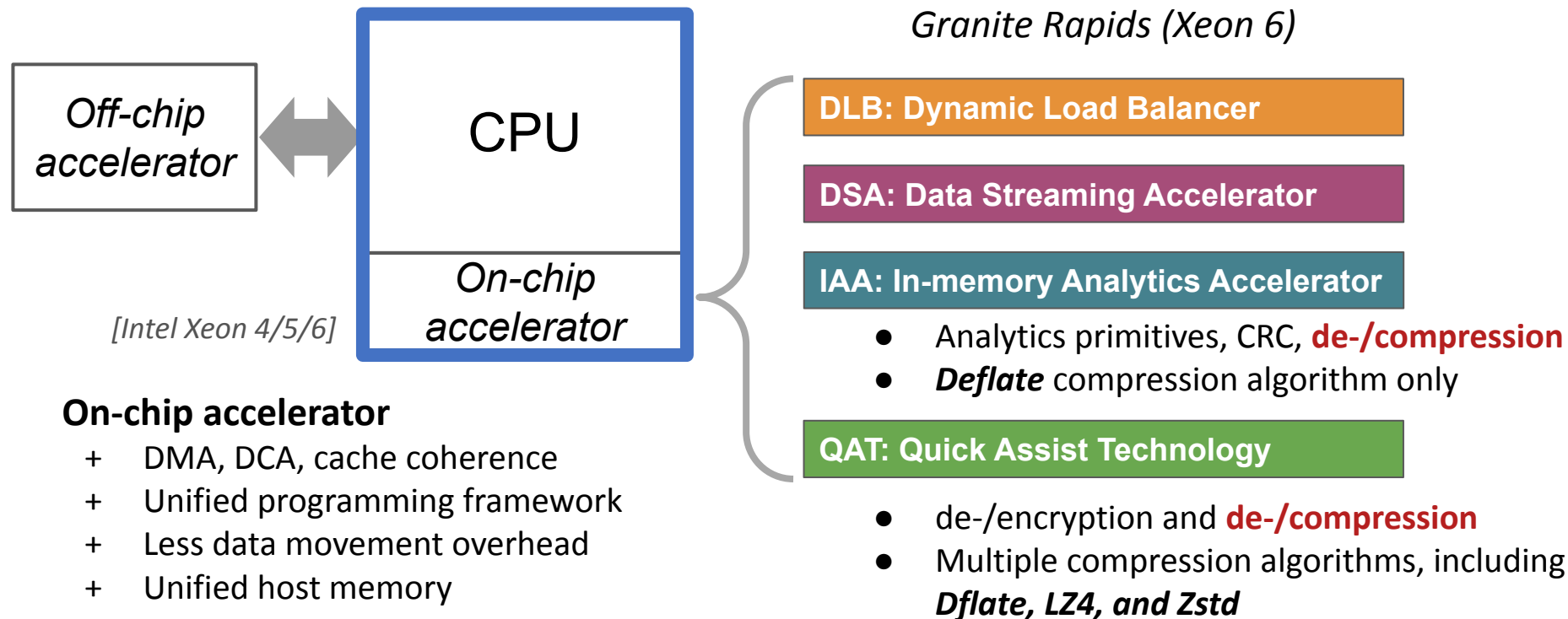
De-/compression: 14-31% of the Datacenter tax

*[Profiling Hyperscale Big Data Processing, ISCA'23]*

- Additional critical path component
- Wasted CPU cycles
- LLC & memBW contention

## Accelerate de-/compression to reduce Datacenter tax

# Modern CPUs enable multiple decompression paths



**Data systems need to decide where to de-/compress**

# Experimental setup

## Compression algorithms

- Deflate: static & dynamic Huffman coding
- LZ4
- Zstd (compression only)

## Dataset

- TPC-H @SF=10 lineitem, binary columnar format

## Hardware

CPU: Intel Xeon 6952P CPU @ 2.1GHz

- 2x 96 cores, hyper-threading disabled
- Cache: L1: 21 MiB; L2: 384 MiB; LLC: 960 MiB
- Memory: 1 TiB
- *lzbench 2.1.1*

On-chip QAT: 4 /socket

- *Intel QATzip 1.3.2, QAT-ZSTD-PLUGIN 1.0.0*

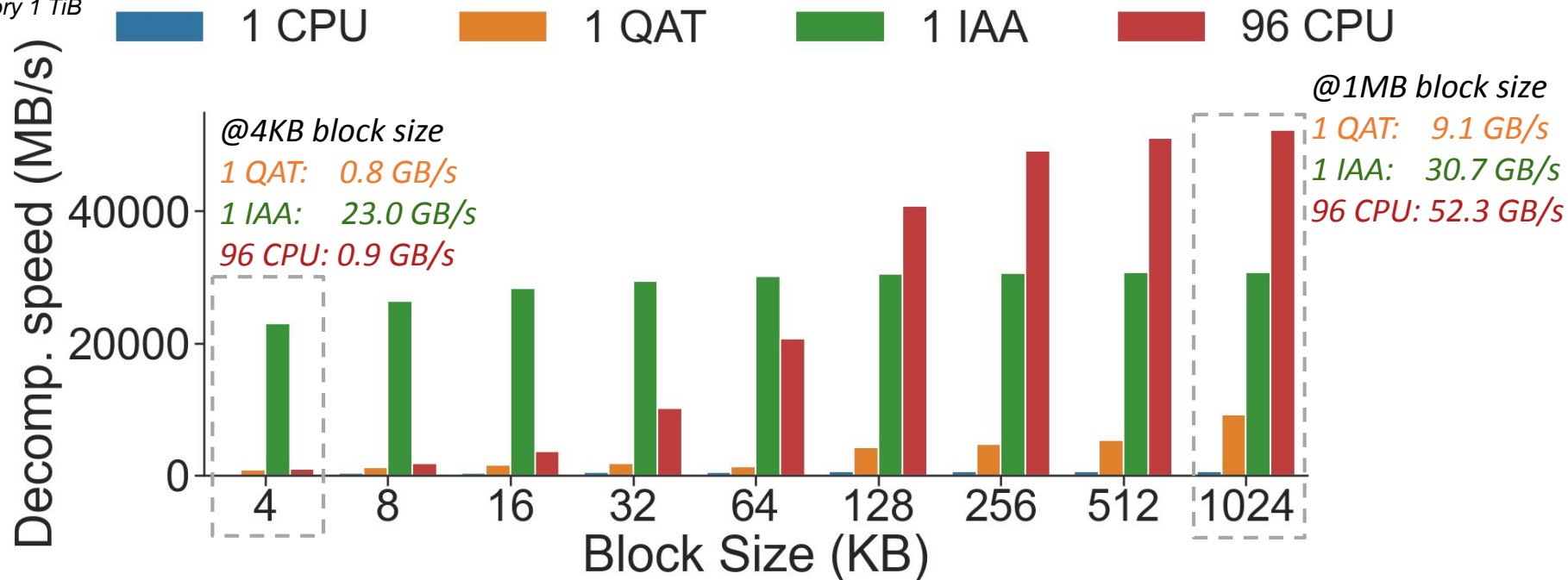
On-chip IAA: 4 /socket, deflate only

- *Intel QPL 1.9.0*

# Block size determines the faster execution path

CPU: Intel 6952P @ 2.1GHz locked  
192 cores, hyper-threading disabled  
L1: 21 MiB, L2: 384 MiB, LLC: 960 MiB  
Memory 1 TiB

Deflate (dynamic)



Small blocks → IAA; large blocks → QAT / CPU parallelism

# The free lunch has a queue: offload overhead

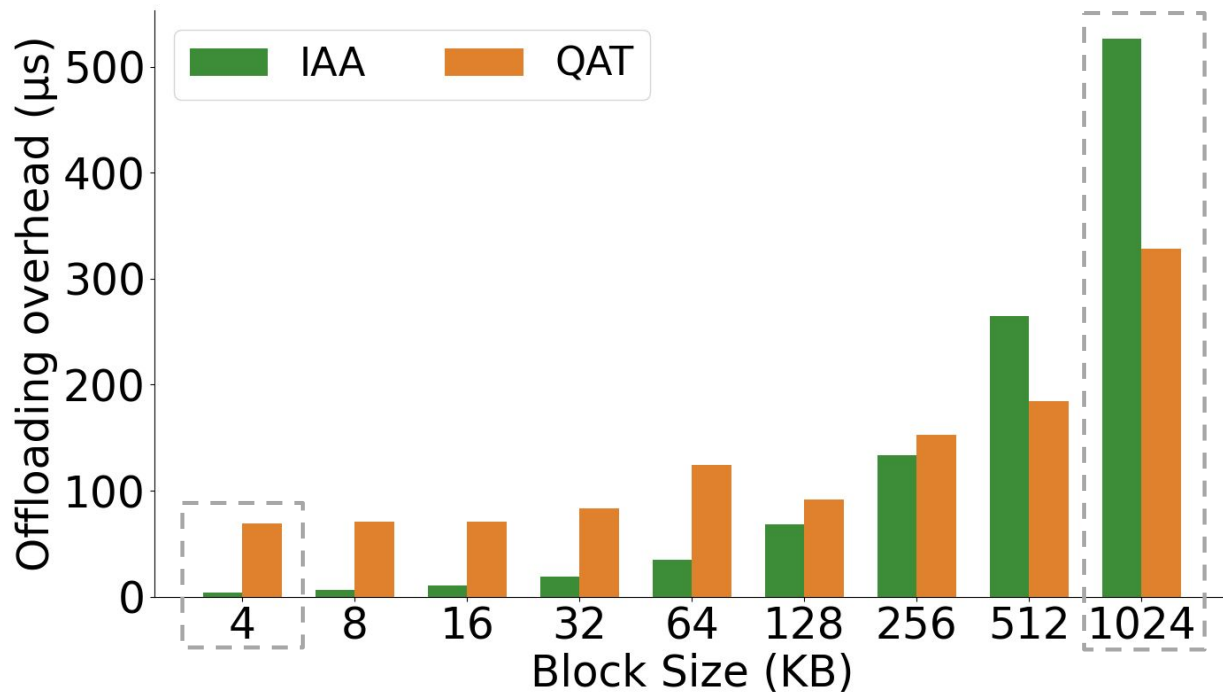
CPU: Intel 6952P @ 2.1GHz locked  
192 cores, hyper-threading disabled  
L1: 21 MiB, L2: 384 MiB, LLC: 960 MiB  
Memory 1 TiB

$$T_{accel} =$$

$$T_{submit} + T_{HW\_execution} + T_{poll}$$

Offloading overhead

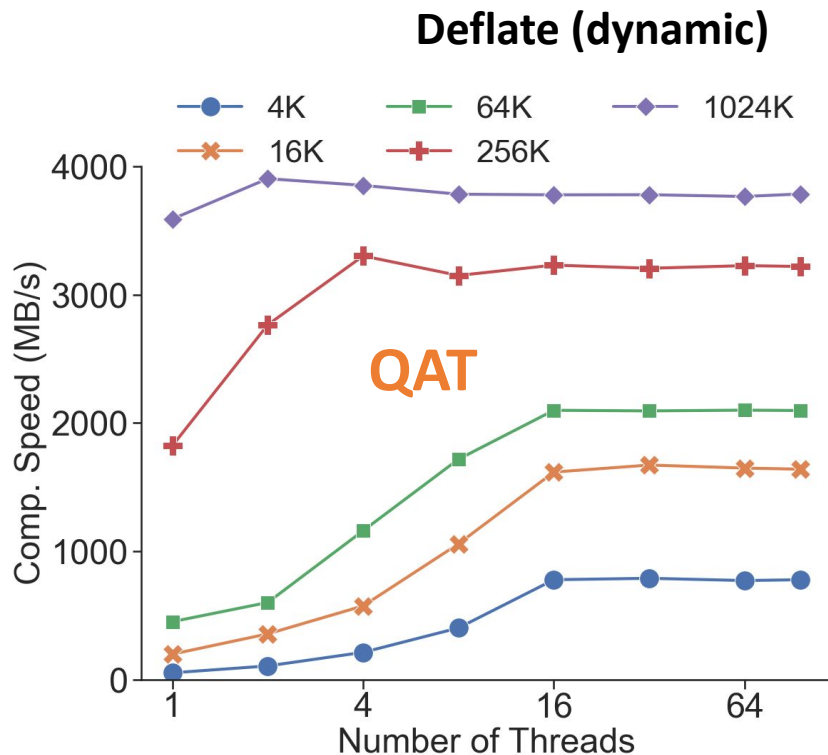
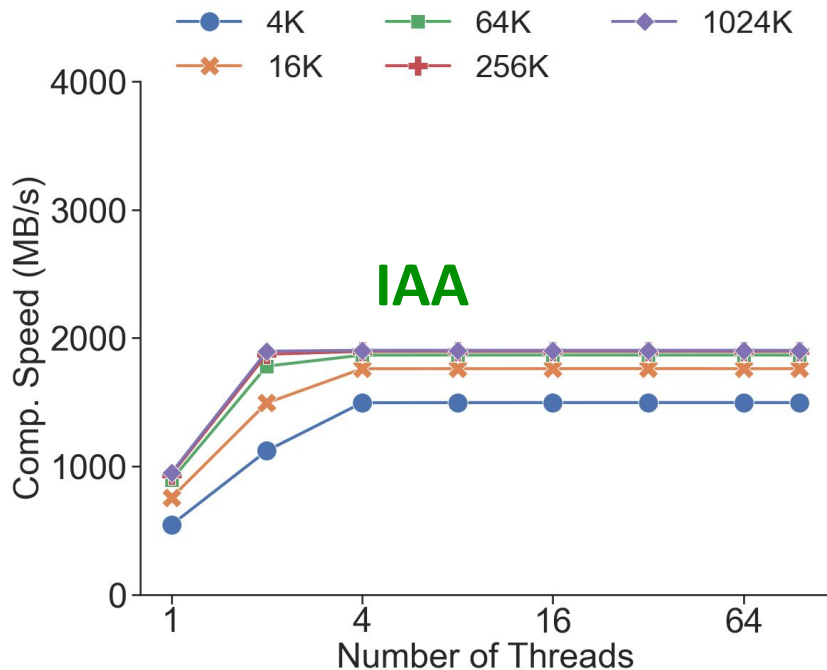
- Format descriptor
- Issue *ENQCMD* instruction
- Wait for completion



**Offload granularity must match the accelerator**

# Saturating the on-chip accelerators

**CPU:** Intel 6952P @ 2.1GHz locked  
192 cores, hyper-threading disabled  
L1: 21 MiB, L2: 384 MiB, LLC: 960 MiB  
Memory 1 TiB

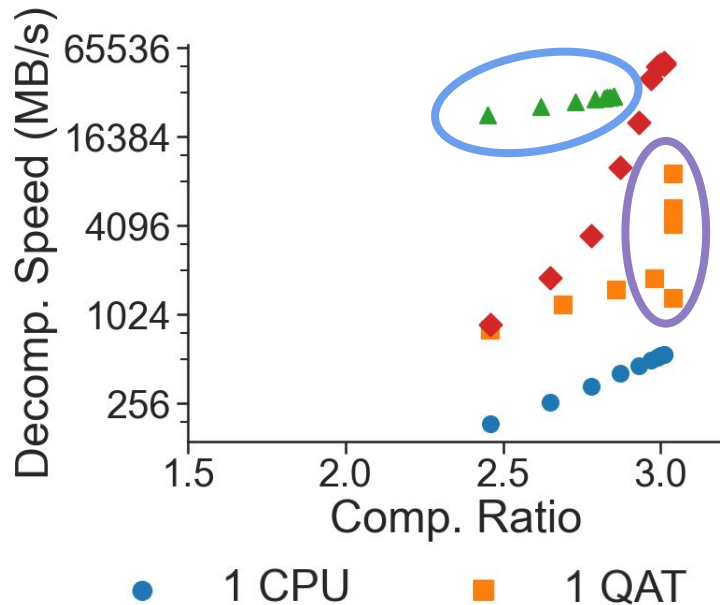


**IAA saturates early; QAT requires more parallel submissions**

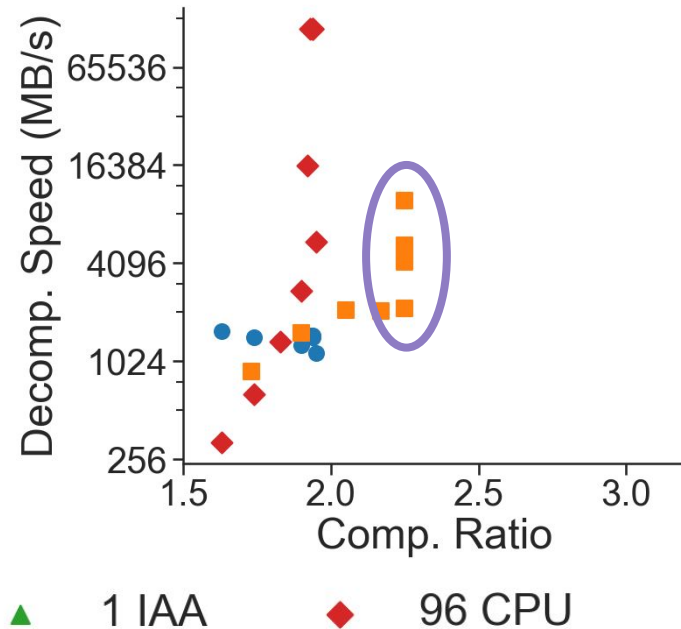
# Choosing the optimal accelerator / algorithm

**CPU:** Intel 6952P @ 2.1GHz locked  
192 cores, hyper-threading disabled  
L1: 21 MiB, L2: 384 MiB, LLC: 960 MiB  
Memory 1 TiB

**Deflate (dynamic)**



**LZ4**



**IAA: Deflate specialization; QAT: multi-codec coverage**

# Analytics must offload de-/compression adaptively

- **Route by codec, operation, block size granularity, and device load**
  - IAA: a strong path for Deflate, including small blocks
  - QAT: provides broader codec coverage
  - CPU: remains competitive for large-block compression
- **Maximize aggregate decompression throughput: IAA-QAT-CPU**

## Future directions

- + Hardware offloading within the context of a complete query engine
- + Dynamic runtime routing (CPU, on-chip & off-chip accelerators)

**Thank you!**