Hunting for Offload: Automated Discovery of Acceleratable Code in Datacenters

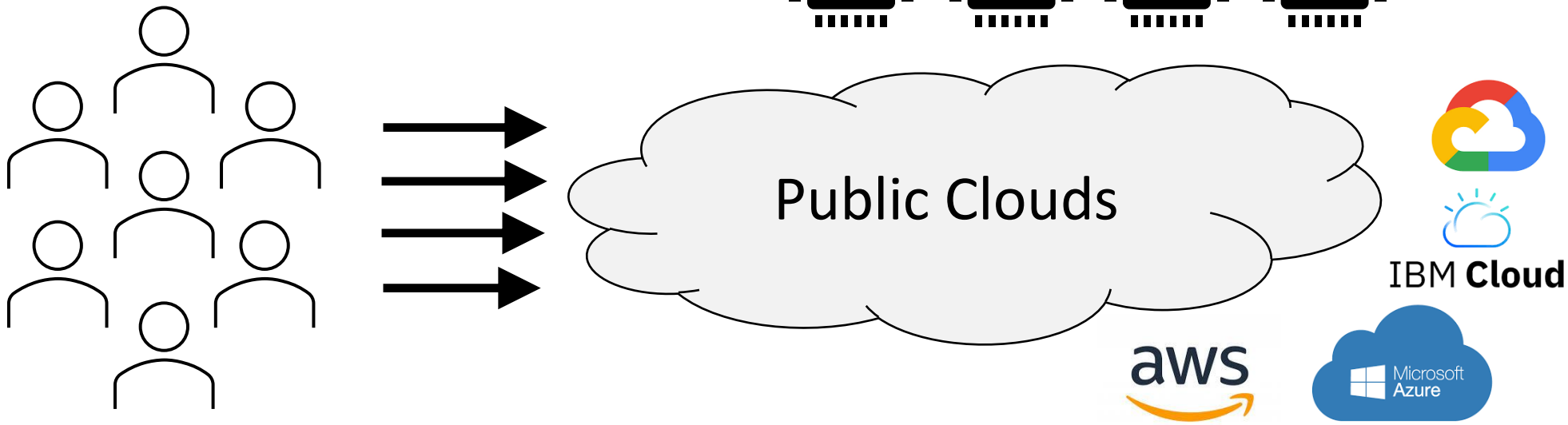
Joshua Kim[1], Chaojie Zhang[2], Íñigo Goiri[2], Christopher J. Rossbach[1,2], Jovan Stojkovic[1]

[1]*The University of Texas at Austin,* [2]*Microsoft*

The Growth of Cloud Computing

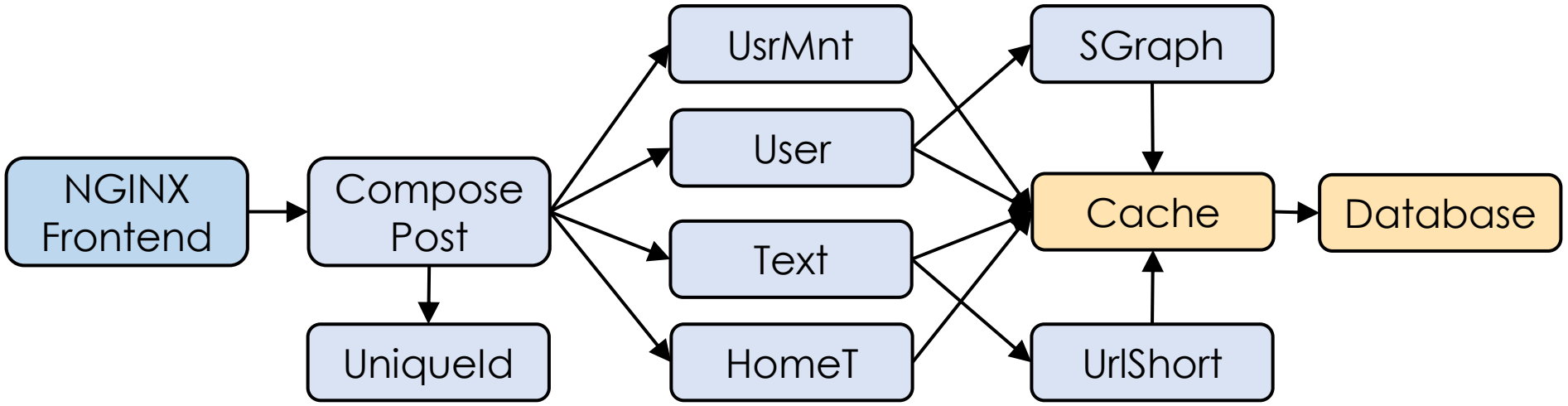
New Computing Paradigms:

- ***Microservices***
- ***Serverless or Function-as-a-Service (FaaS)***



Microservices

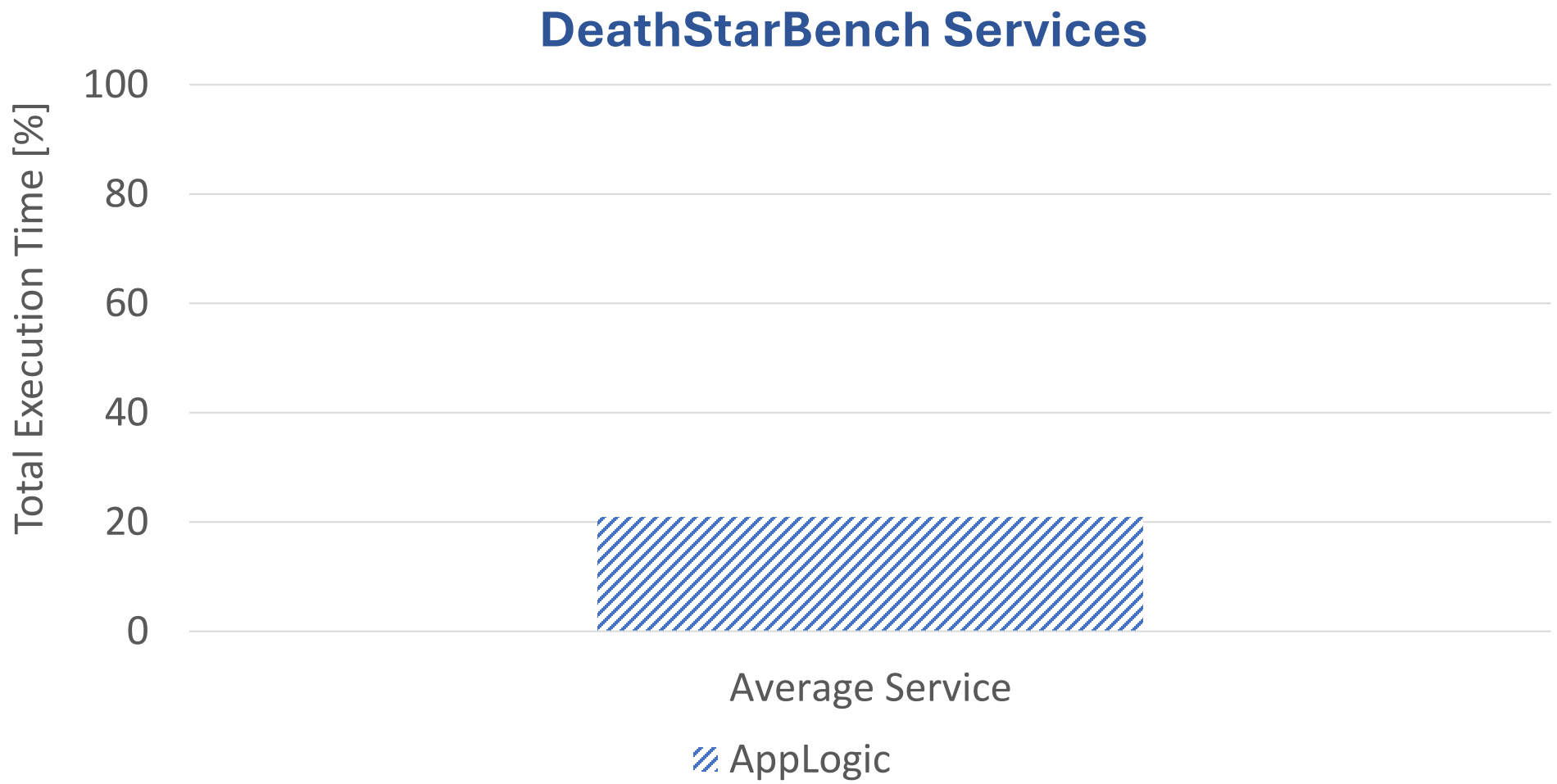
- Large monolithic applications decomposed into many small interdependent services
  - Each service implements separate functionality



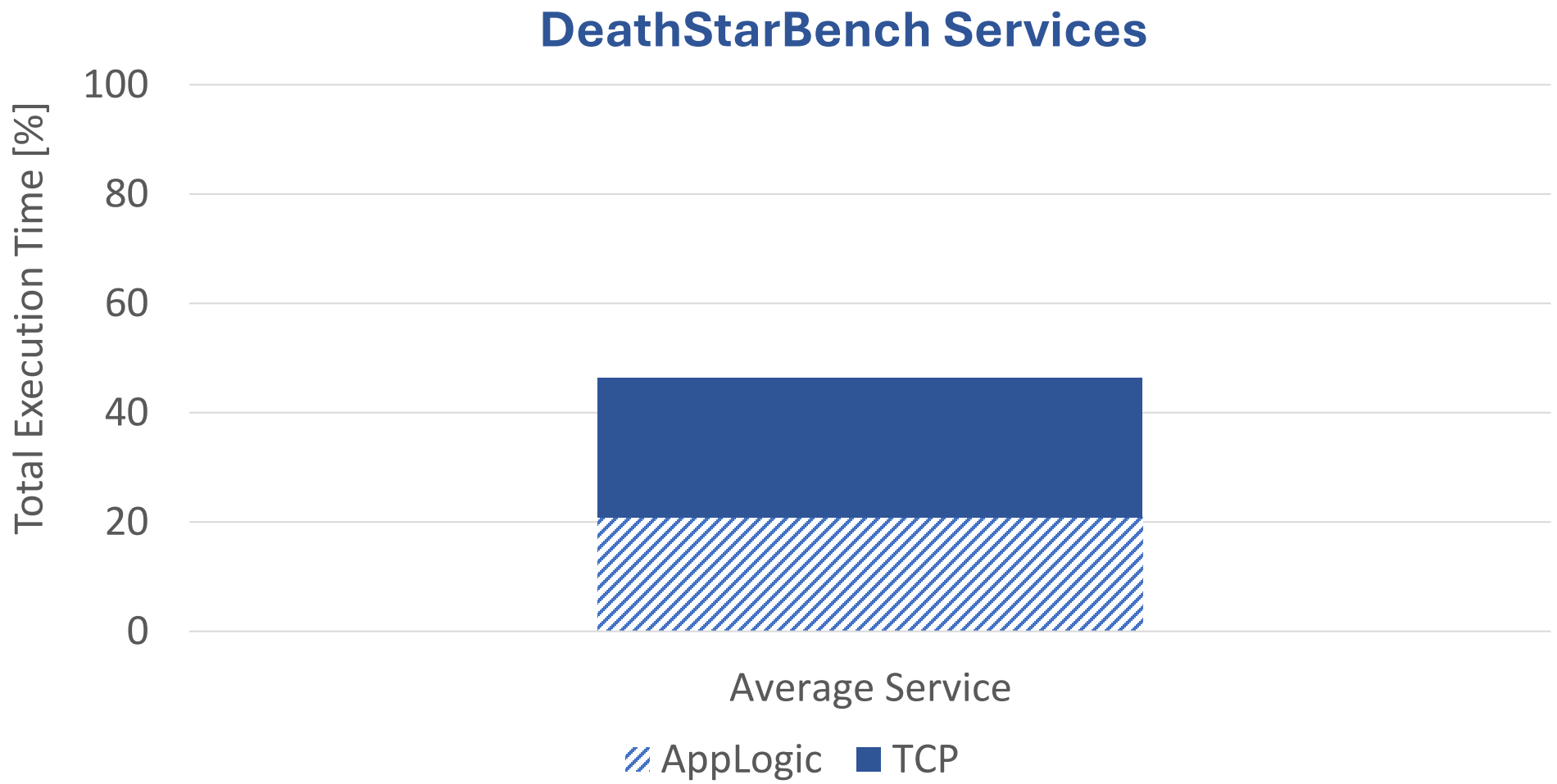
Benefits of Microservices

- Scalability 
- Design simplicity 
- HW management 

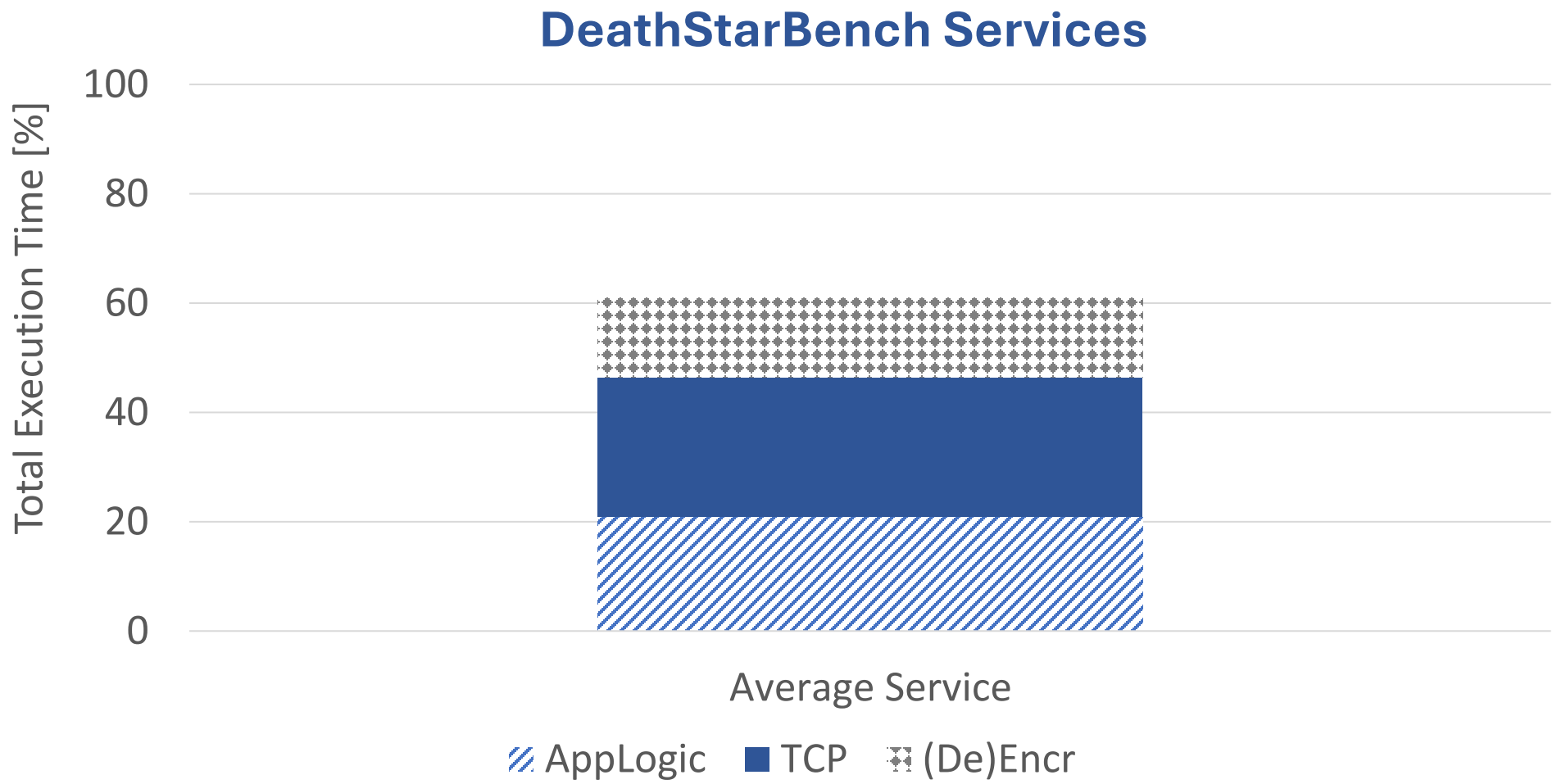
Datacenter Tax Dominates Execution



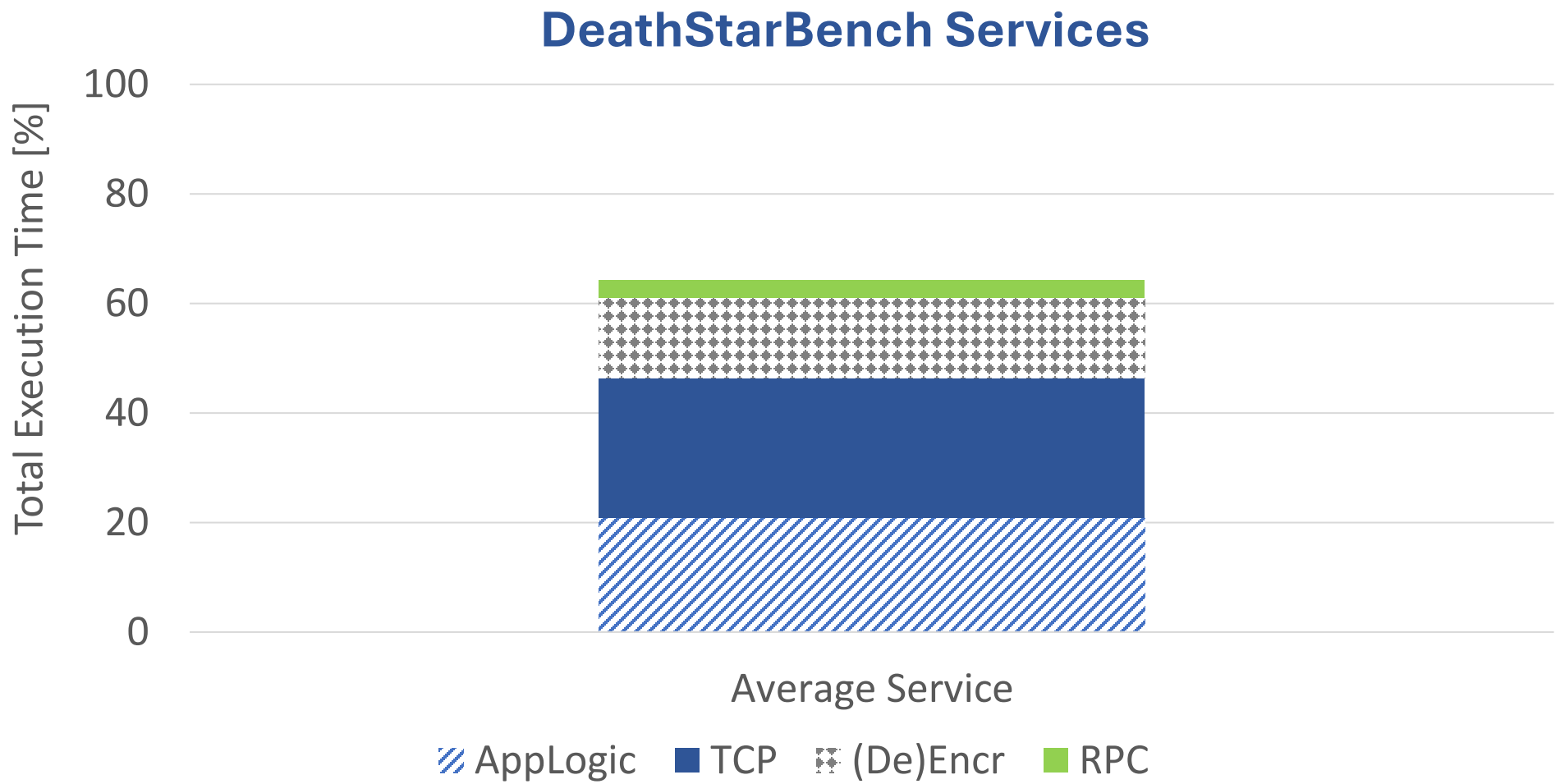
Datacenter Tax Dominates Execution



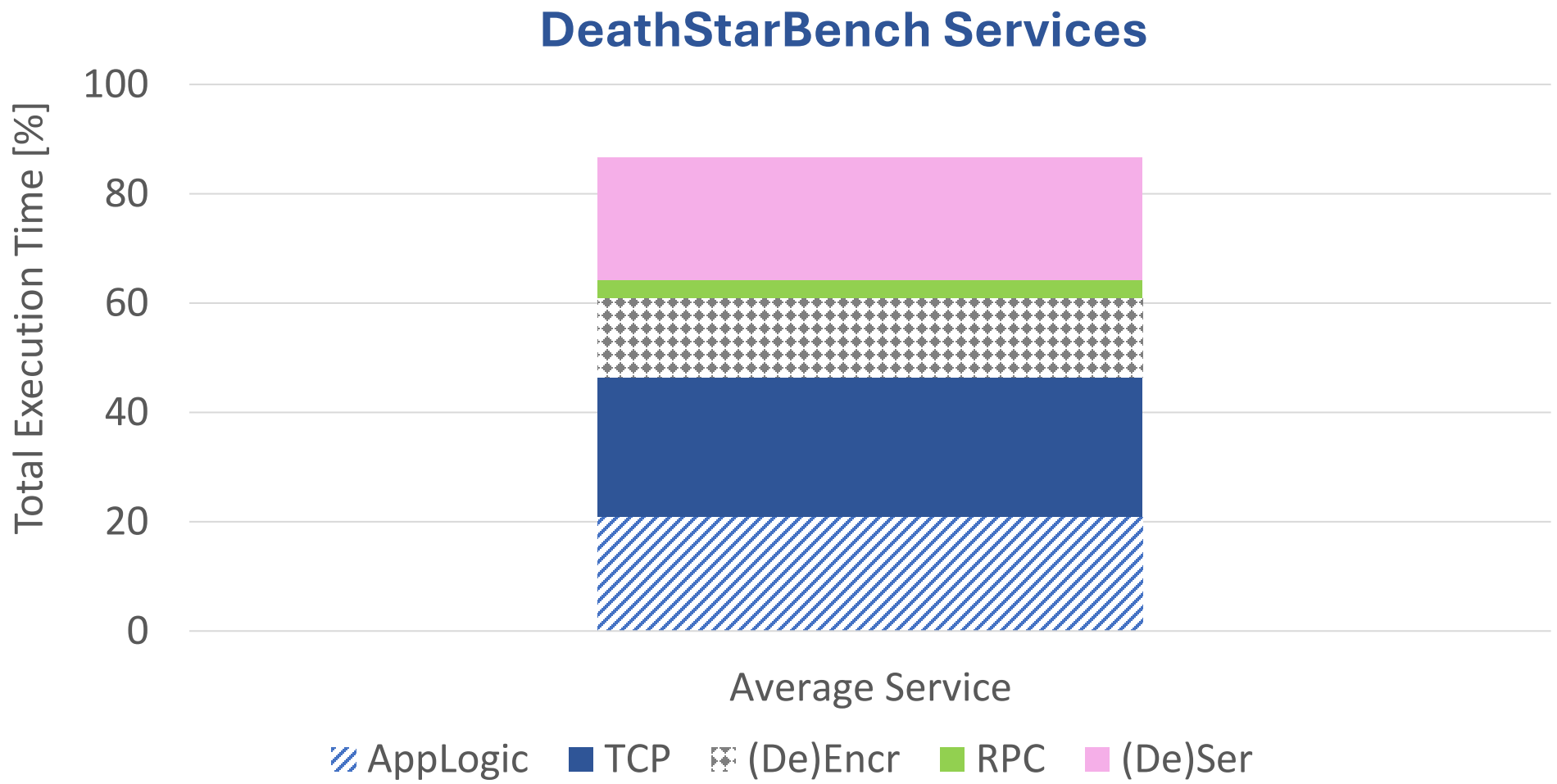
Datacenter Tax Dominates Execution



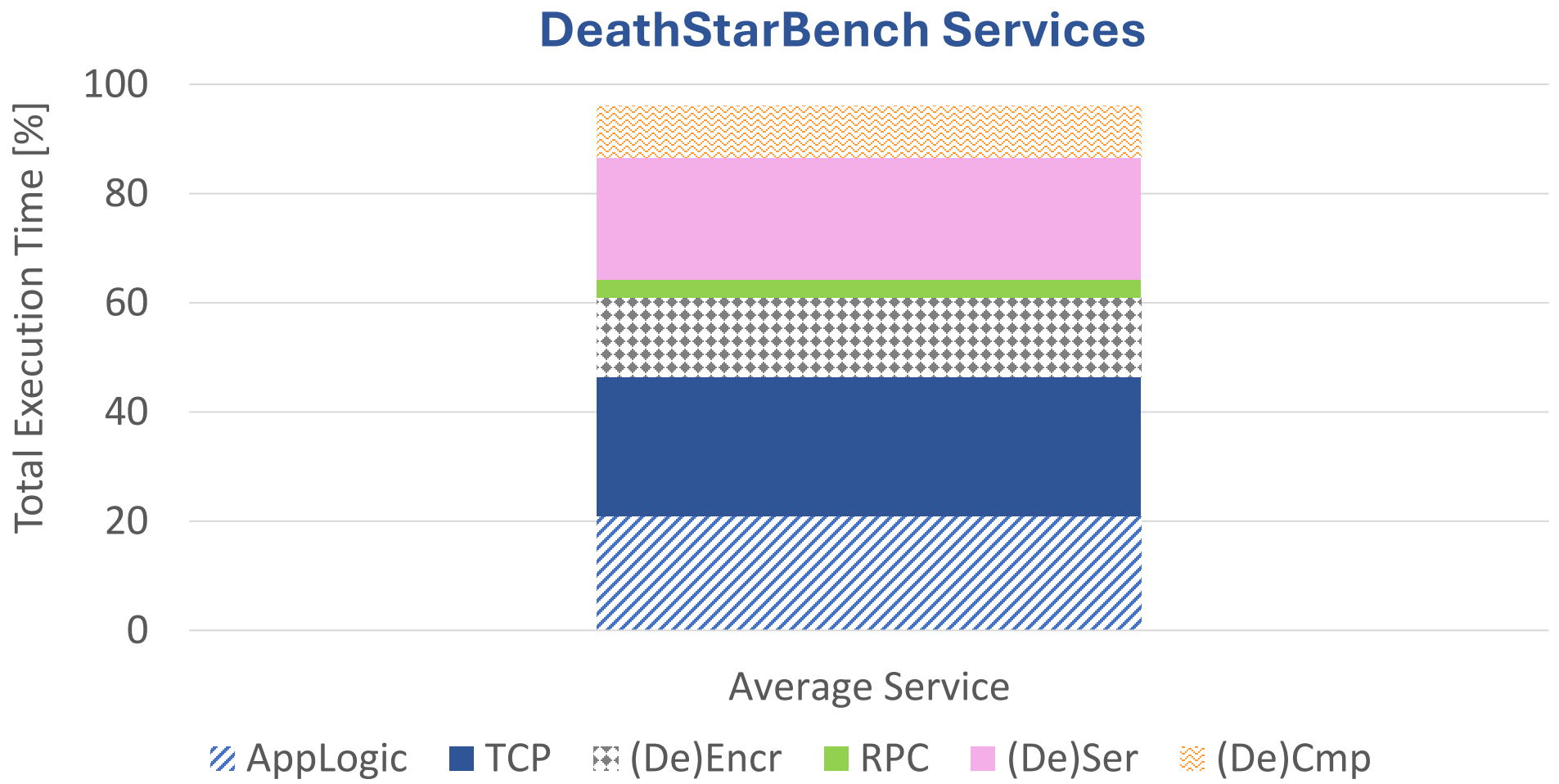
Datacenter Tax Dominates Execution



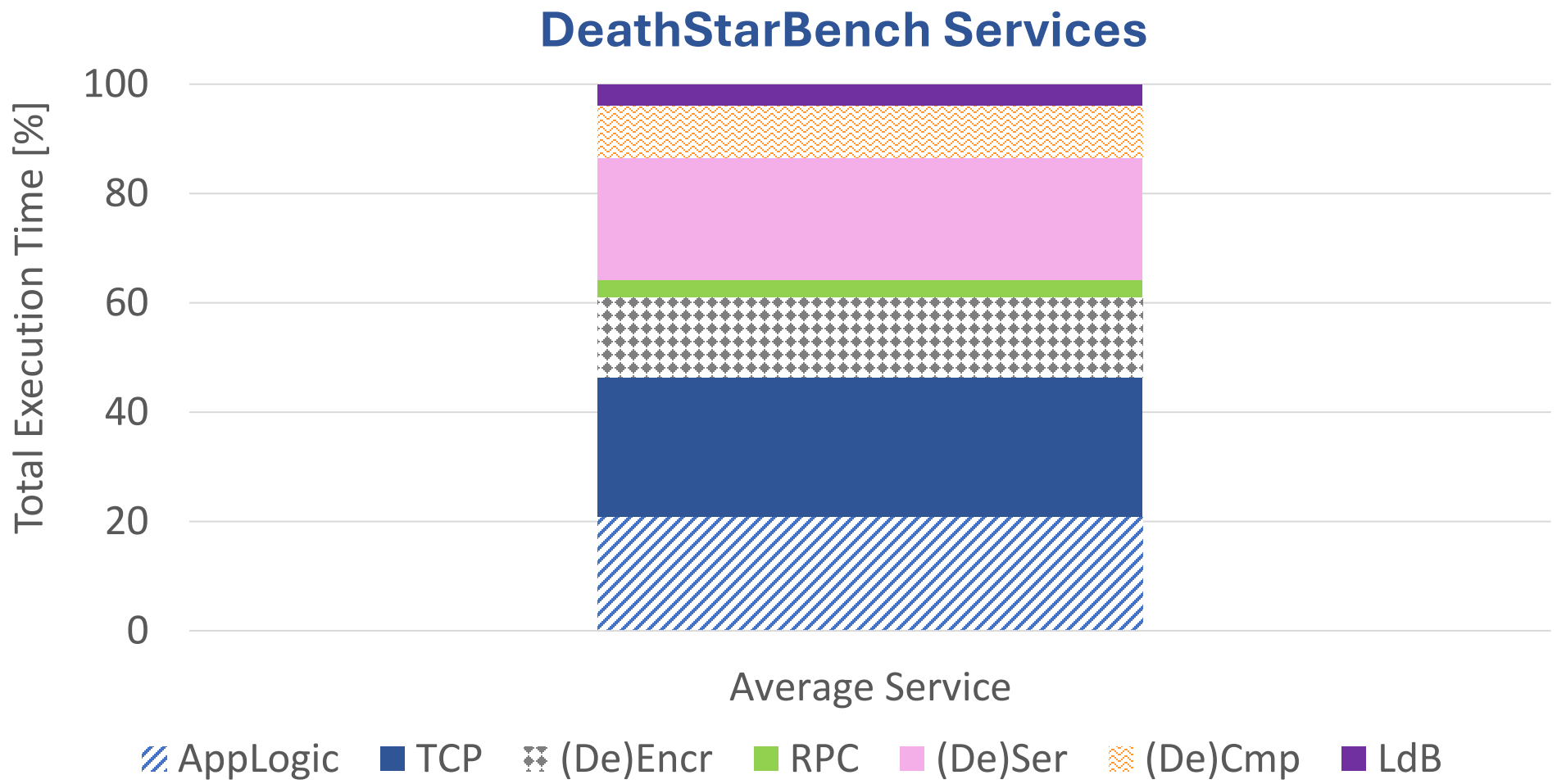
Datacenter Tax Dominates Execution



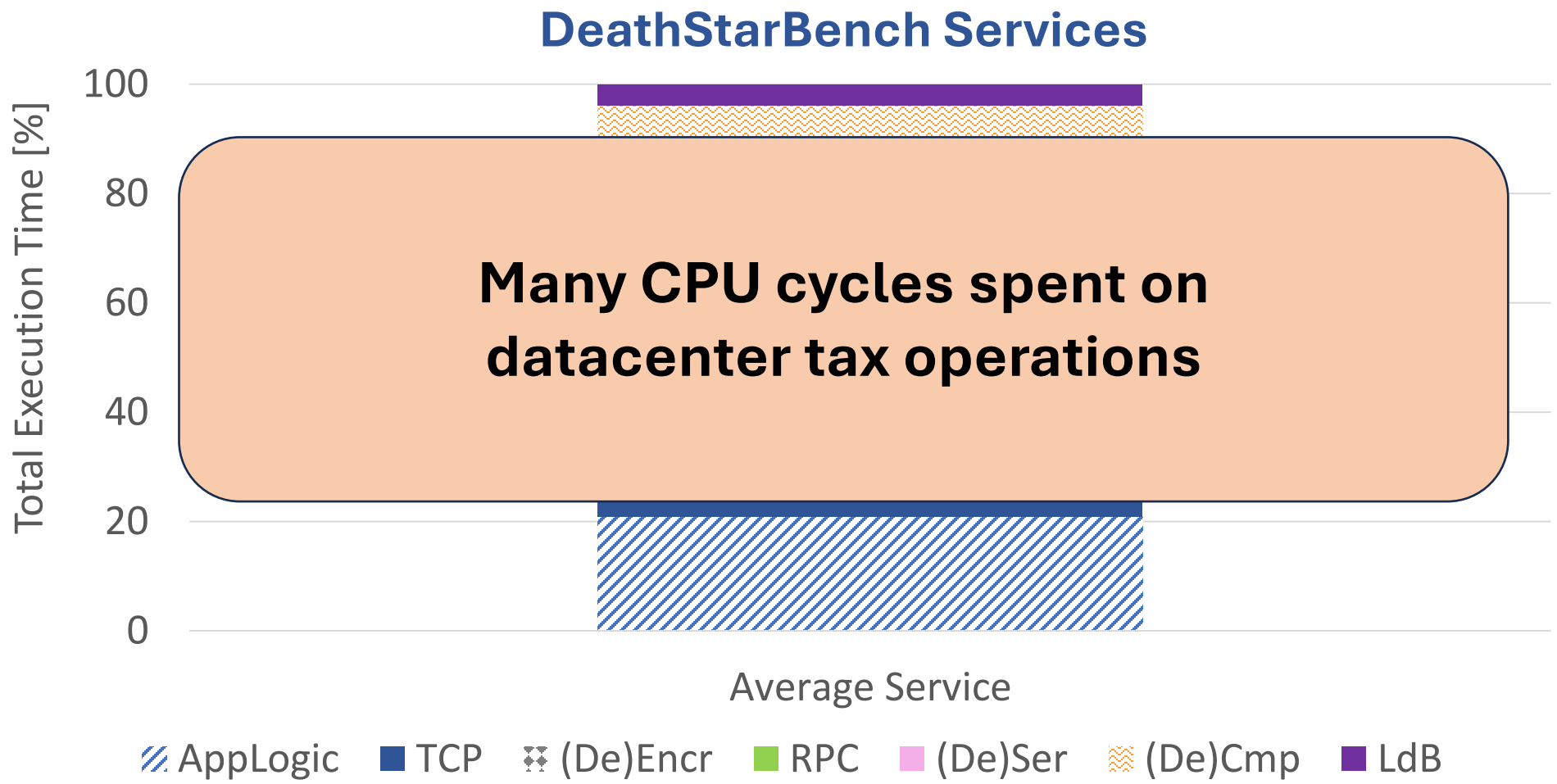
Datacenter Tax Dominates Execution



Datacenter Tax Dominates Execution



Datacenter Tax Dominates Execution



Proposals For Individual Accelerators

CDPU: Co-designing Compression and Decompression Processing Units for Hyperscale Systems

Sagar Karandikar
UC Berkeley, Google
Berkeley, CA, USA

Aniruddha N. Udipi
Google
Mountain View, CA, USA

Junsun Choi
UC Berkeley
Berkeley, CA, USA

Joonho Whangbo
UC Berkeley
Berkeley, CA, USA

Jerry Zhao
UC Berkeley
Berkeley, CA, USA

Svilen Kanev
Google
Mountain View, CA, USA

Edwin Lim
UC Berkeley
Berkeley, CA, USA

Jyrki Alakuijala
Google
Zürich, Switzerland

Vrishab Madduri
UC Berkeley
Berkeley, CA, USA

Yakun Sophia Shao
UC Berkeley
Berkeley, CA, USA

Borivoje Nikolić
UC Berkeley
Berkeley, CA, USA

Krste Asanović
UC Berkeley
Berkeley, CA, USA

F4T: A Fast and Flexible FPGA-based Full-stack TCP Acceleration Framework

Junehyuk Boo
junehyuk@snu.ac.kr
Seoul National University
MangoBoost Inc.
Seoul, Republic of Korea

Yujin Chung
yujin.chung@snu.ac.kr
Seoul National University
MangoBoost Inc.
Seoul, Republic of Korea

Eunjin Baek
ebaek@snu.ac.kr
Seoul National University
MangoBoost Inc.
Seoul, Republic of Korea

Seongmin Na
seongmin.na@snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Changsu Kim
changsu.kim@mangoboost.io
MangoBoost Inc.
Seoul, Republic of Korea

Jangwoo Kim*
jangwoo@snu.ac.kr
Seoul National University
MangoBoost Inc.
Seoul, Republic of Korea

Dagger: Efficient and Fast RPCs in Cloud Microservices with Near-Memory Reconfigurable NICs

Nikita Lazarev
Cornell University
Ithaca, New York, USA
nl524@cornell.edu

Shaojie Xiang
Cornell University
Ithaca, New York, USA
sx233@cornell.edu

Neil Adit
Cornell University
Ithaca, New York, USA
na469@cornell.edu

Zhiru Zhang
Cornell University
Ithaca, New York, USA
zhiruz@cornell.edu

Christina Delimitrou
Cornell University
Ithaca, New York, USA
delimitrou@cornell.edu

A Hardware Accelerator for Protocol Buffers

Sagar Karandikar
UC Berkeley, Google
USA

Chris Leary
Google
USA

Chris Kennelly
Google
USA

Jerry Zhao
UC Berkeley
USA

Dinesh Parimi
UC Berkeley
USA

Borivoje Nikolić
UC Berkeley
USA

Krste Asanović
UC Berkeley
USA

Parthasarathy Ranganathan
Google
USA

Proposals For Individual Accelerators

CDPU: Co-designing Compression and Decompression Processing Units for Hyperscale Systems

Sagar Karandikar
UC Berkeley, Google
Berkeley, CA, USA

Joonho Whangbo
UC Berkeley
Berkeley, CA, USA

Edwin Lim
UC Berkeley
Berkeley, CA, USA

Yakun Sophia Shao
UC Berkeley

F4T: A Fast and Fl… TCP Acce…

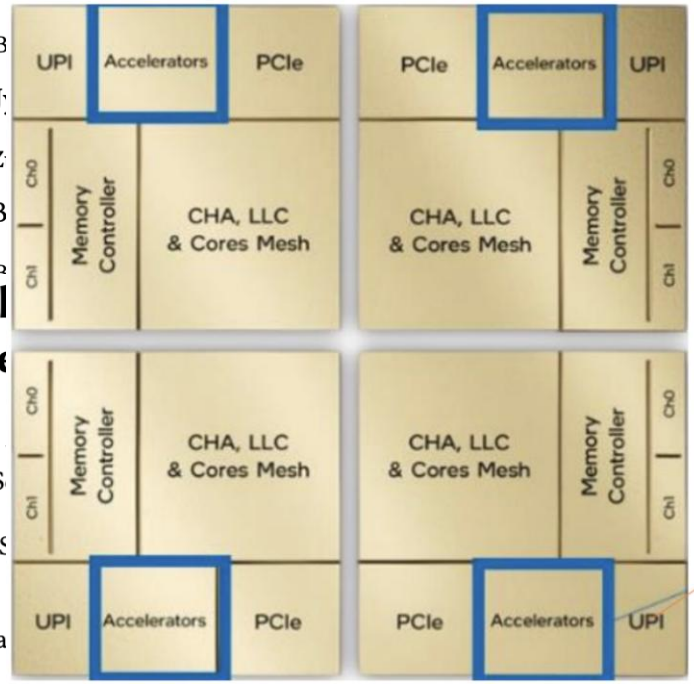
Junehyuk Boo
junehyuk@snu.ac.kr
Seoul National University
MangoBoost Inc.
Seoul, Republic of Korea

Seongmin Na
seongmin.na@snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Intel's SoC

Aniruddha N. U…
Google
Mountain View, CA…

Jerry Zhao
Svilen Kanev



Seoul, Republic of Korea
MangoBoost Inc.
Seoul, Republic of Korea

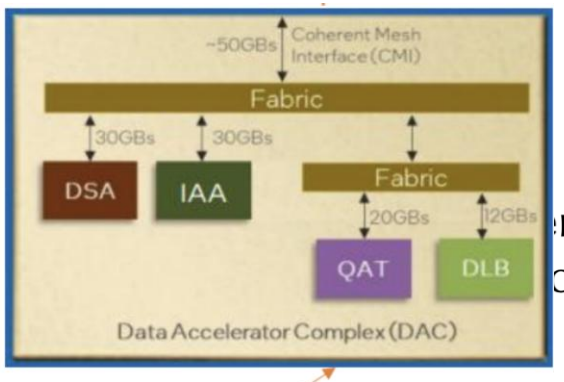
Dagger: Efficient and Fast RPCs in Cloud Microservices with Near-Memory Reconfigurable NICs

nl524@cornell.edu

Shaojie Xiang
Cornell University
Ithaca, New York, USA
sx233@cornell.edu

Neil Adit
Cornell University
Ithaca, New York, USA
na469@cornell.edu

Christina Delimitrou
Cornell University
Ithaca, New York, USA
delimitrou@cornell.edu



Data Accelerator Complex

…rator for Protocol Buffers

Chris Leary
Google
USA

Chris Kennelly
Google
USA

…inesh Parimi
UC Berkeley
USA

Borivoje Nikolić
UC Berkeley
USA

Parthasarathy Ranganathan
Google
USA

Proposals For Individual Accelerators

CDPU: Co-designing Compression and Decompression Processing Units for Hyperscale Systems

Sagar Karandikar
UC Berkeley, Google
Berkeley, CA, USA

Joonho Whangbo
UC Berkeley
Berkeley, CA, USA

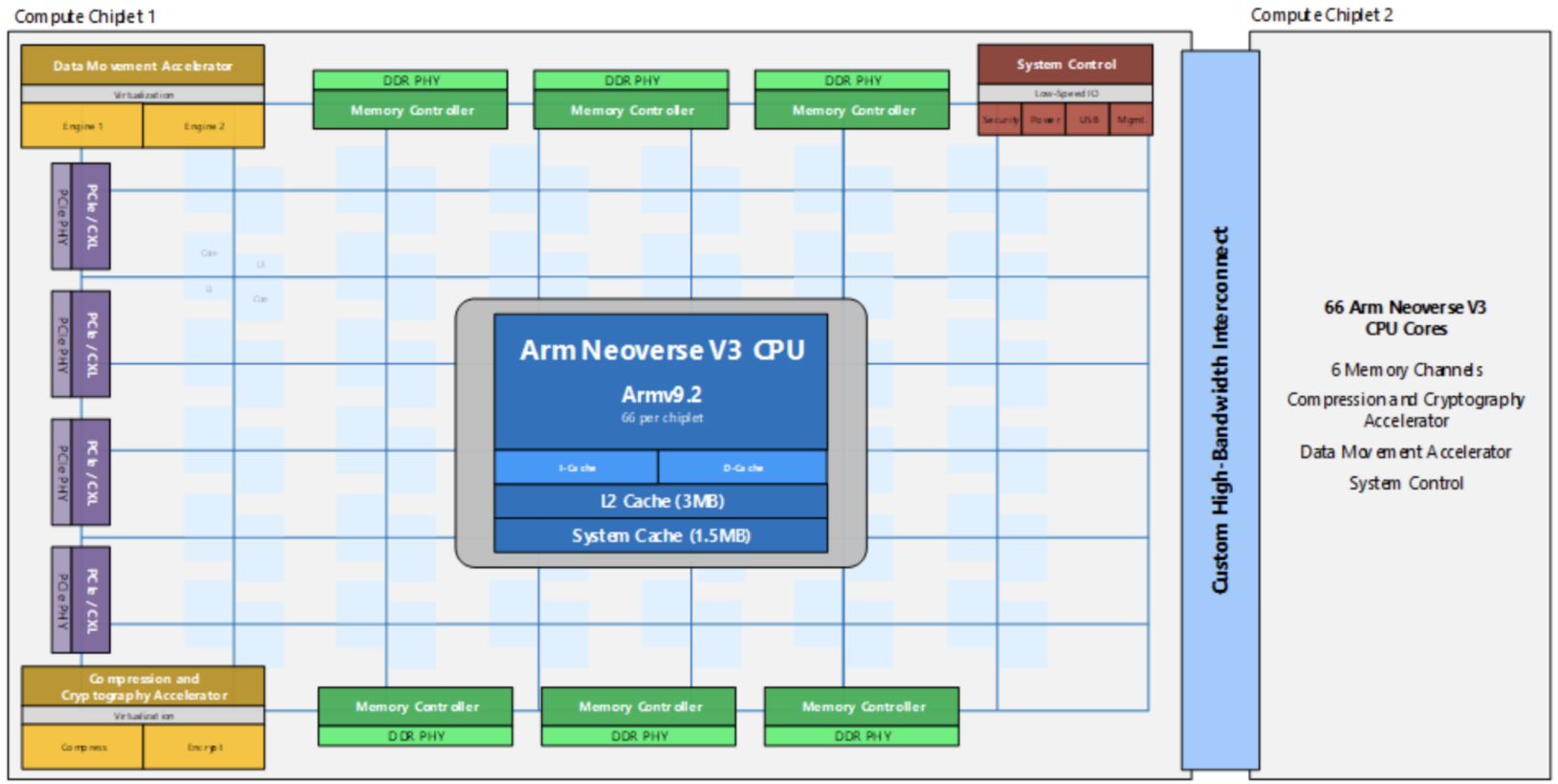
Edwin Lim
UC Berkeley
Berkeley, CA, USA

Yakun Sophia Shao
UC Berkeley
Berkeley, CA, USA

F4T: A F

Junehyuk Boo
junehyuk@snu.ac.kr
Seoul National Universit
MangoBoost Inc.
Seoul, Republic of Kore

Seongmin Na
seongmin.na@snu.ac.kr
Seoul National University
Seoul, Republic of Korea



Dagger: Efficient and Fast RPCs in Cloud Microservices with Near-Memory Reconfigurable NICs

Neil Adit
Cornell University
Ithaca, New York, USA
na469@cornell.edu

ristina Delimitrou
Cornell University
aca, New York, USA
mitrou@cornell.edu

Protocol Buffers

Chris Kennelly
Google
USA

Borivoje Nikolić
UC Berkeley
USA

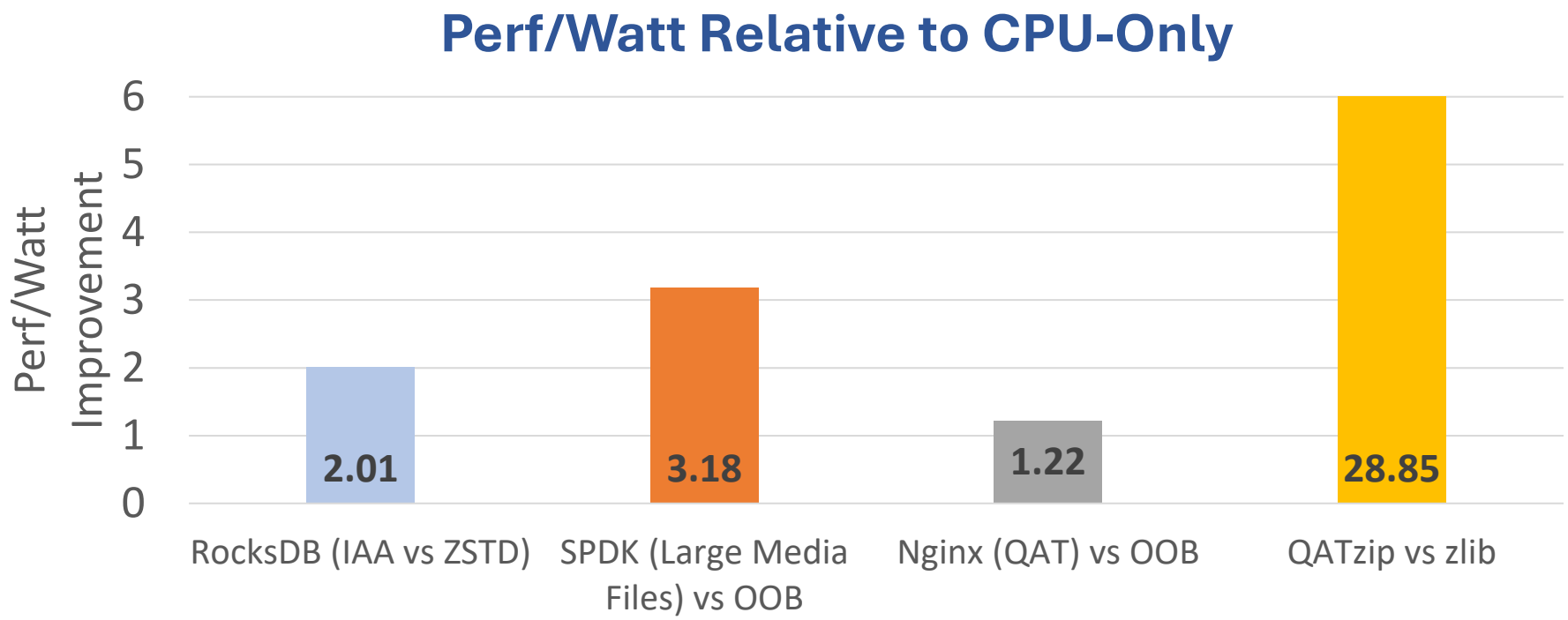
asarathy Ranganathan
Google
USA

cha

Seoul, Republic of Korea

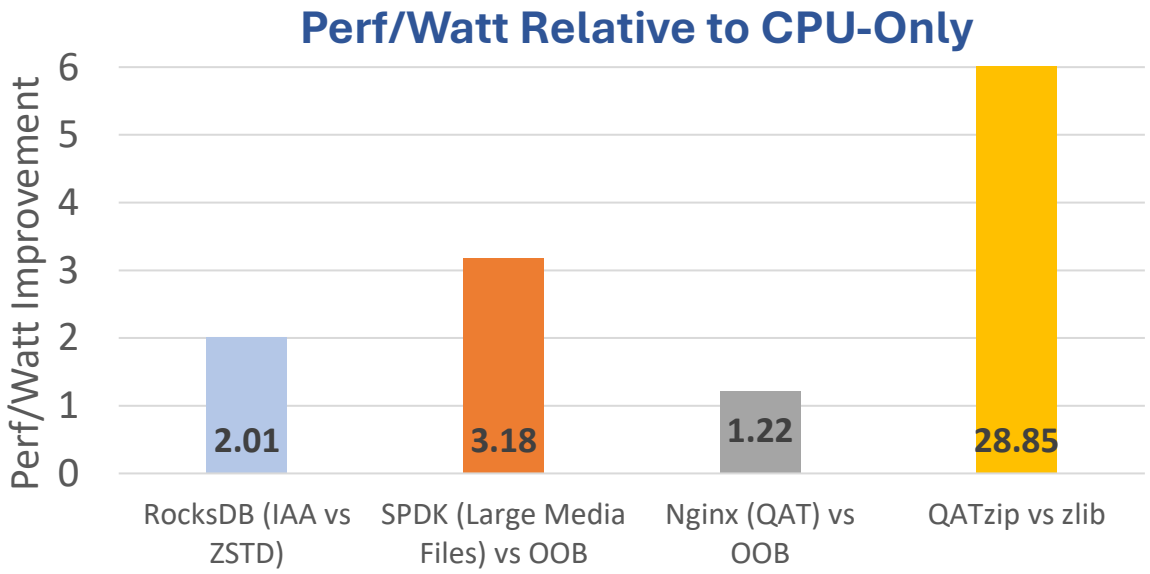
MangoBoost Inc.
Seoul, Republic of Korea

On-Chip Accelerators Unlock Benefits



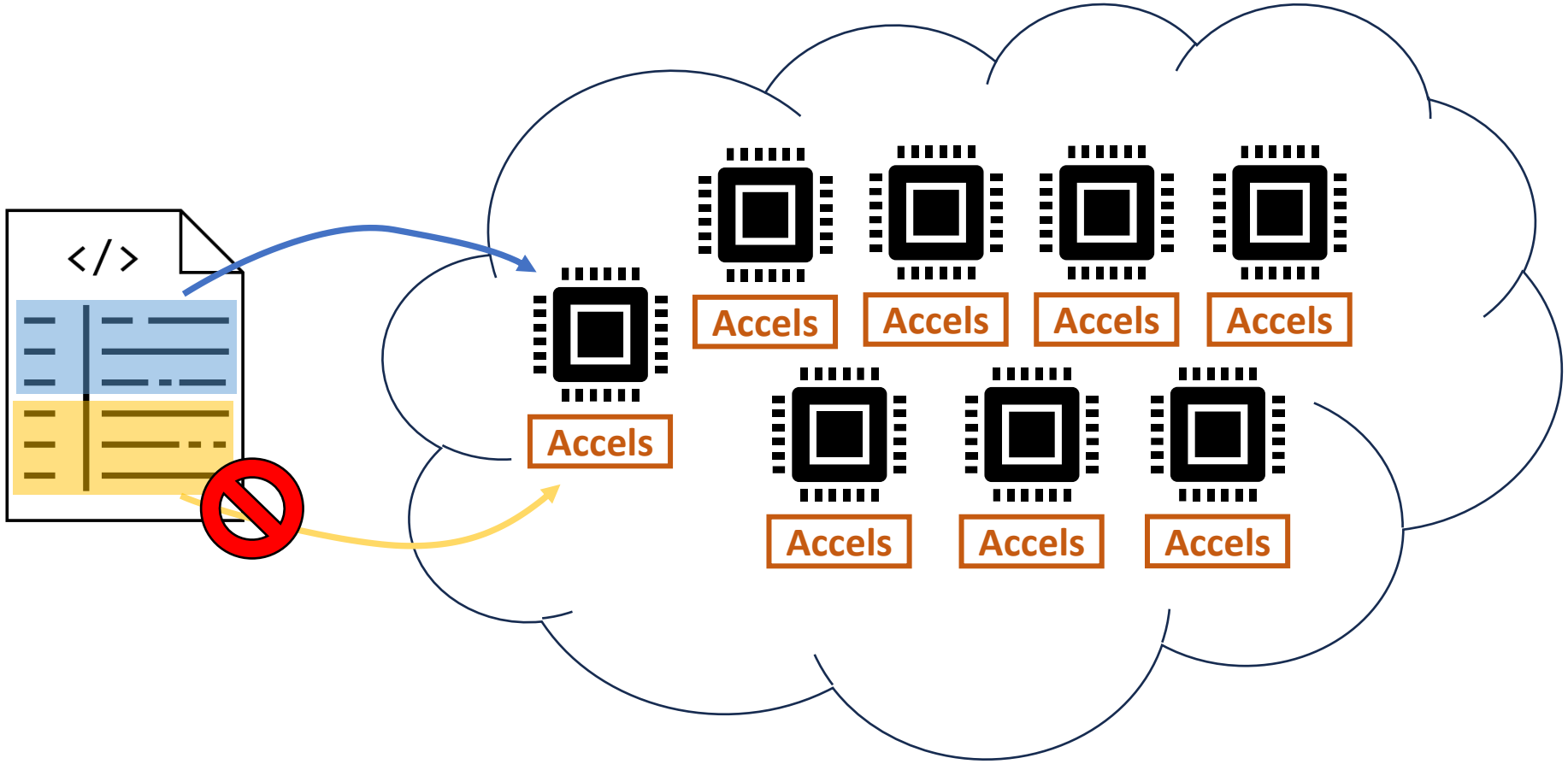
On-Chip Accelerators Unlock Benefits

- Required servers ⇩
- Fleet Power ⇩
- CO2 Emissions ⇩

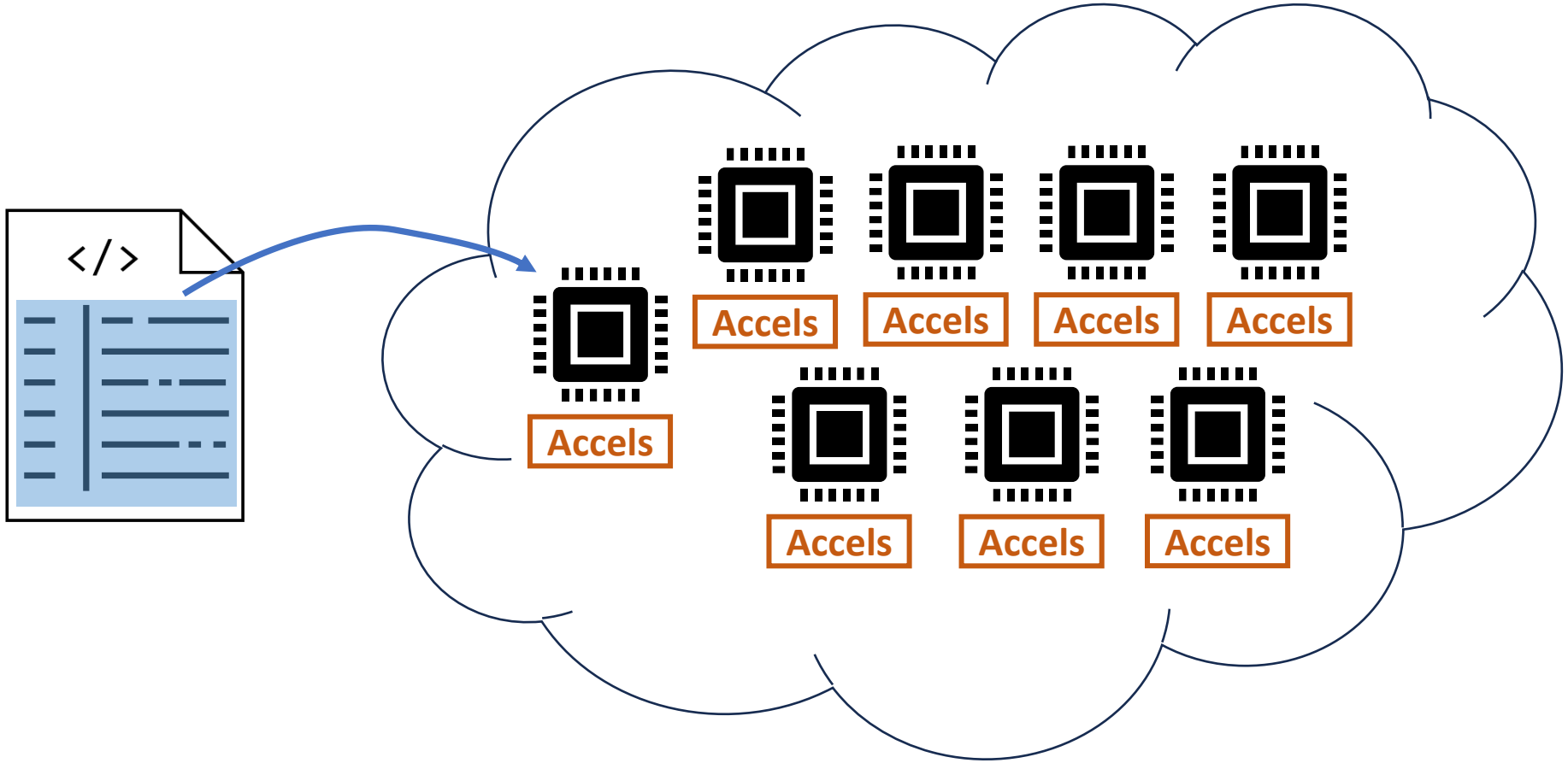


52% - 66% lower TCO across applications

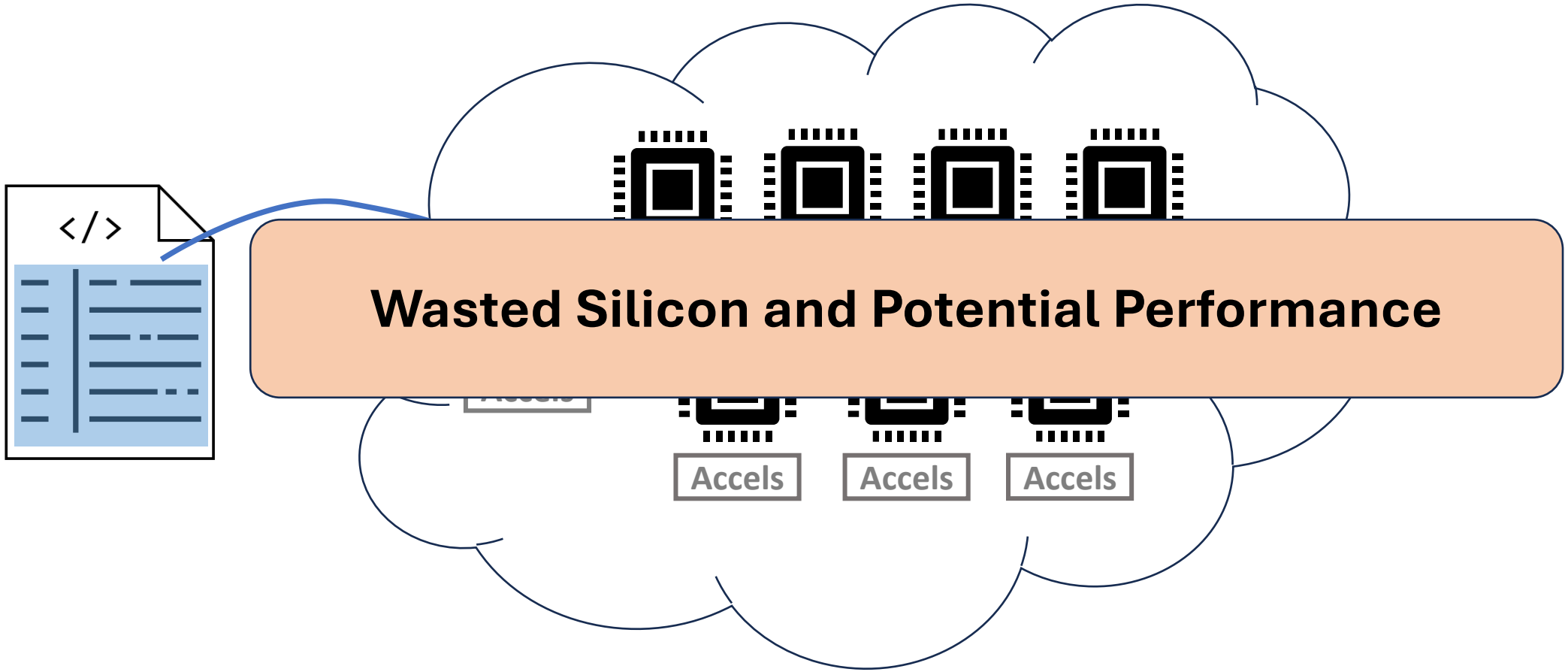
Gap Between Accelerators and Production



Gap Between Accelerators and Production



Gap Between Accelerators and Production



Challenges in Using On-chip Accelerators

- Kernel Drivers Add Complexity

Challenges in Using On-chip Accelerators

- Kernel Drivers Add Complexity
- Different Interfaces

Accel 1 (QAT)

```
sess = cpaDcInitSession (...);
cpaDcCompressData(sess, ...);
```

High-level API Calls

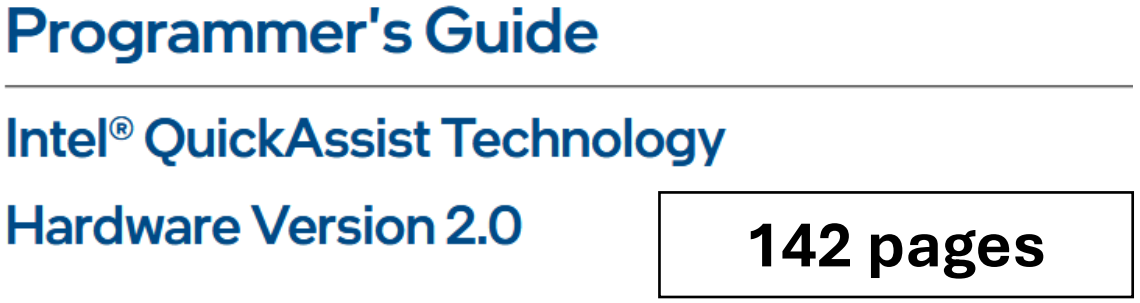
Accel 2 (DSA)

```
struct dsa_hw_desc desc = {0};
desc.opcode = DSA_OPCODE_MEMMOVE;
desc.src_addr = (uintptr_t)src;
                ...
// ENQCMD + Explicit Polling
```

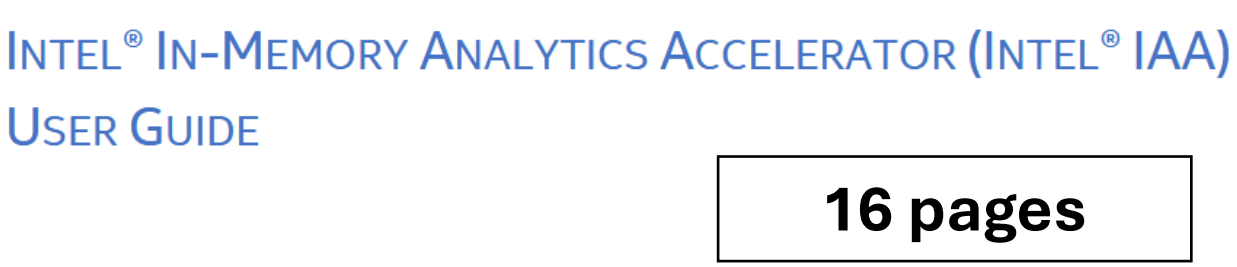
Low-level Descriptor Setup

Challenges in Using On-chip Accelerators

- Kernel Drivers Add Complexity
- Different Interfaces
- High-Barrier to Learning



142 pages



16 pages

"Sea of information"
Steep onboarding curve

Limited options
Requires Implicit Knowledge

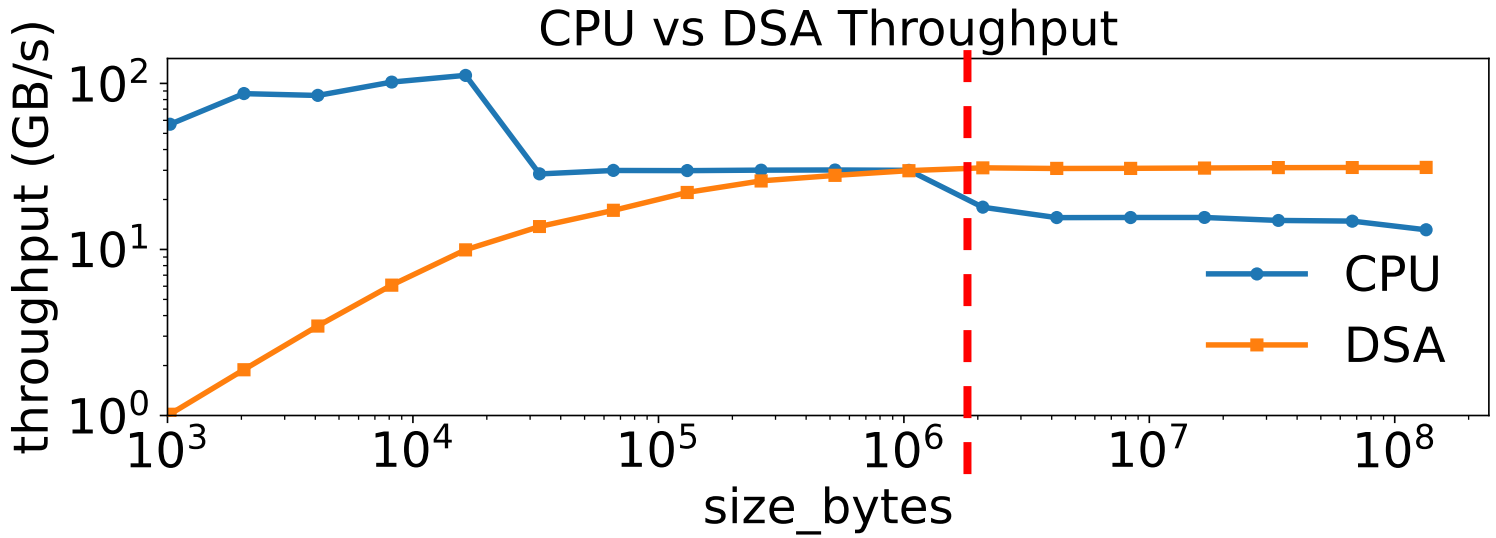
Challenges in EFFICIENT use of Accelerators

- Code+Runtime Dependencies
  - ❓ Object size
  - ❓ Batch size
  - ❓ Request length
  - ❓ Arithmetic Intensity

Challenges in EFFICIENT use of Accelerators

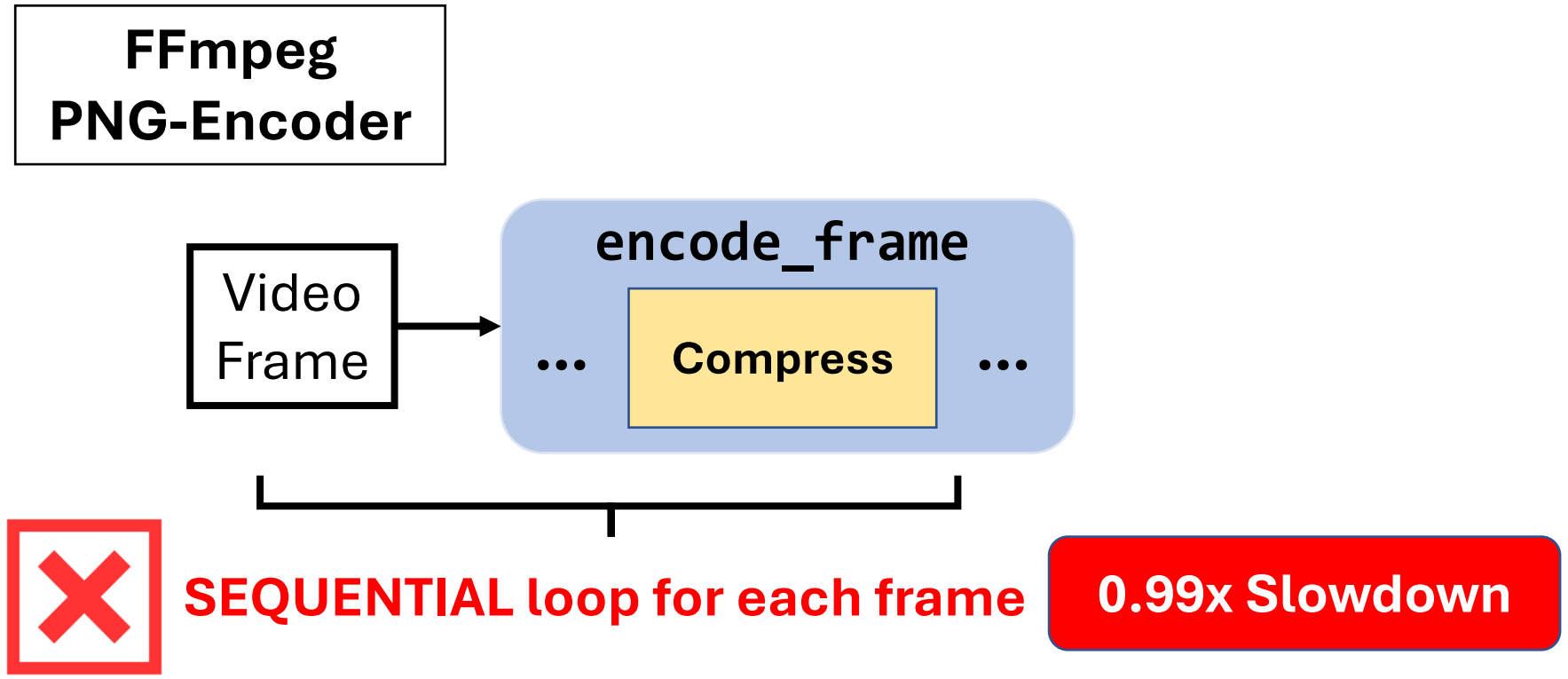
- Code+Runtime Dependencies

- Object size
- Batch size
- Request length
- Arithmetic Intensity



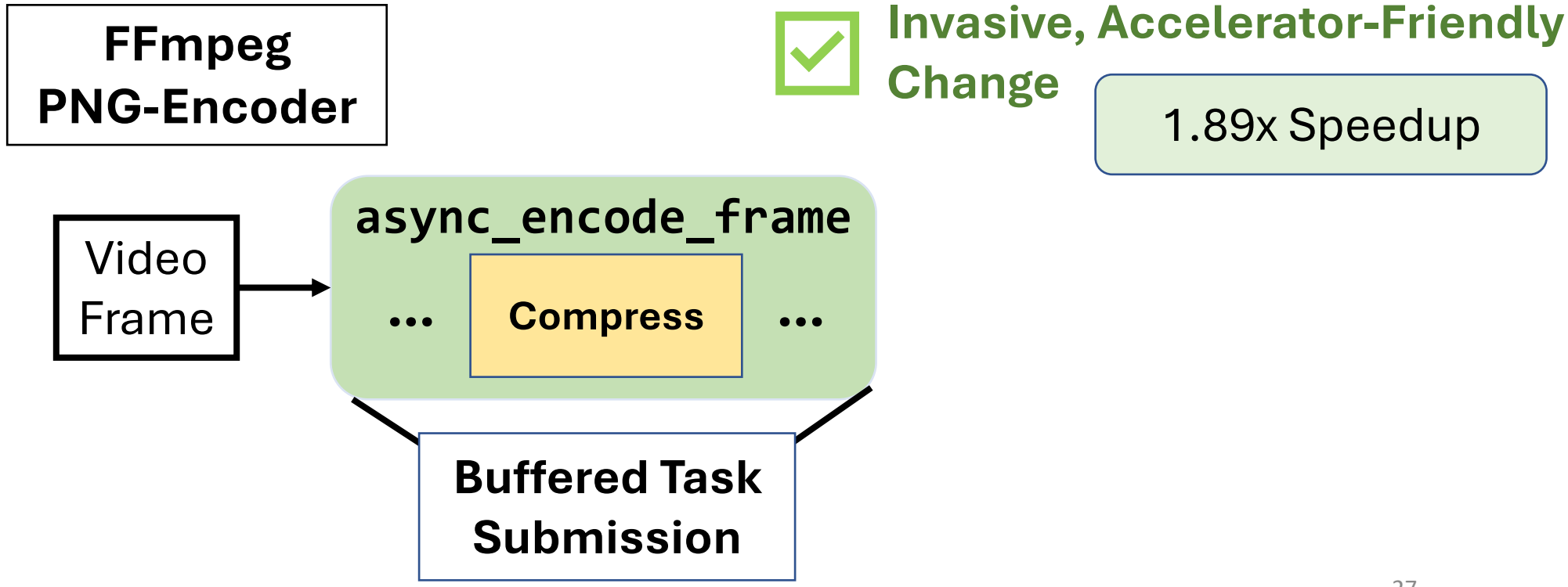
Challenges in EFFICIENT use of Accelerators

- Code+Runtime Dependencies
- Requiring Invasive Re-designs



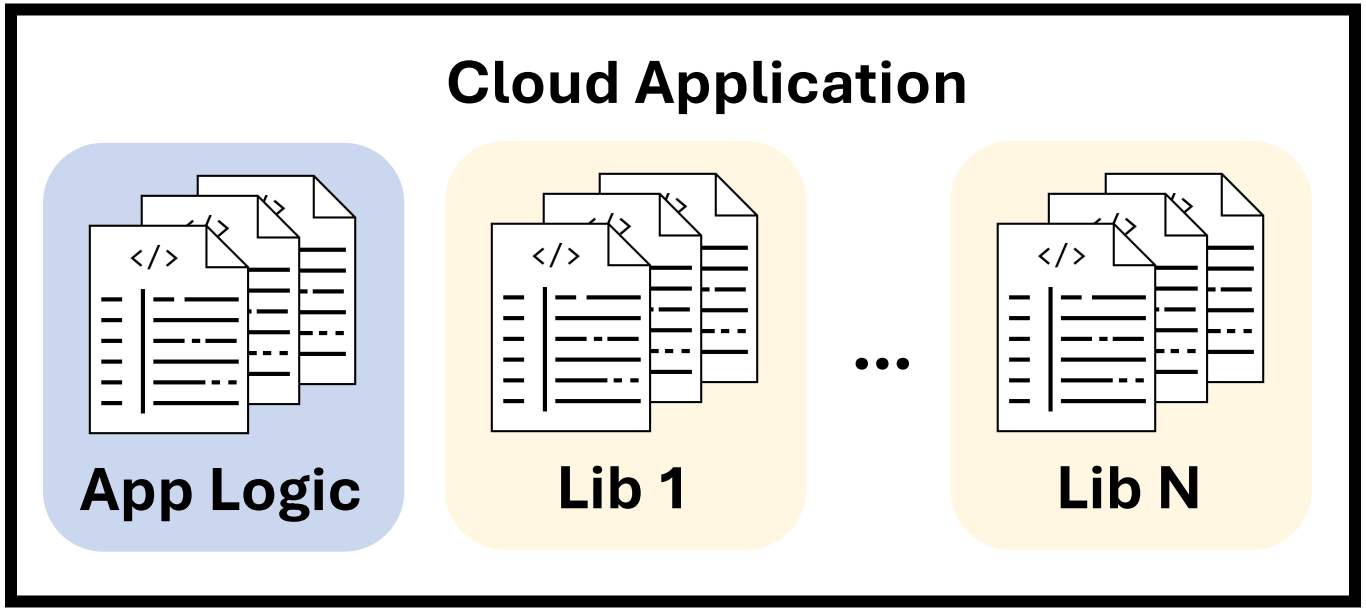
Challenges in EFFICIENT use of Accelerators

- Code+Runtime Dependencies
- Requiring Invasive Re-designs



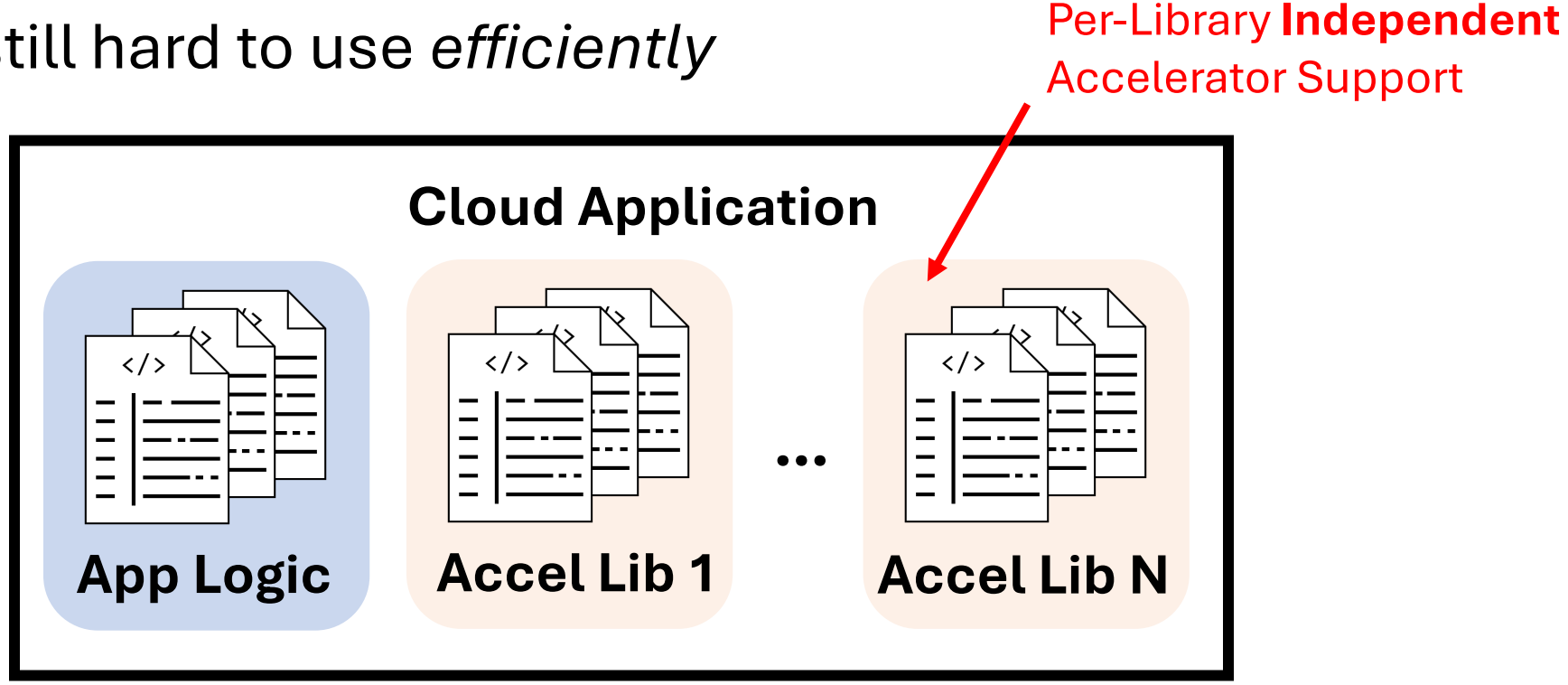
Limitation of SOTA Software Support

- Libraries still hard to use *efficiently*



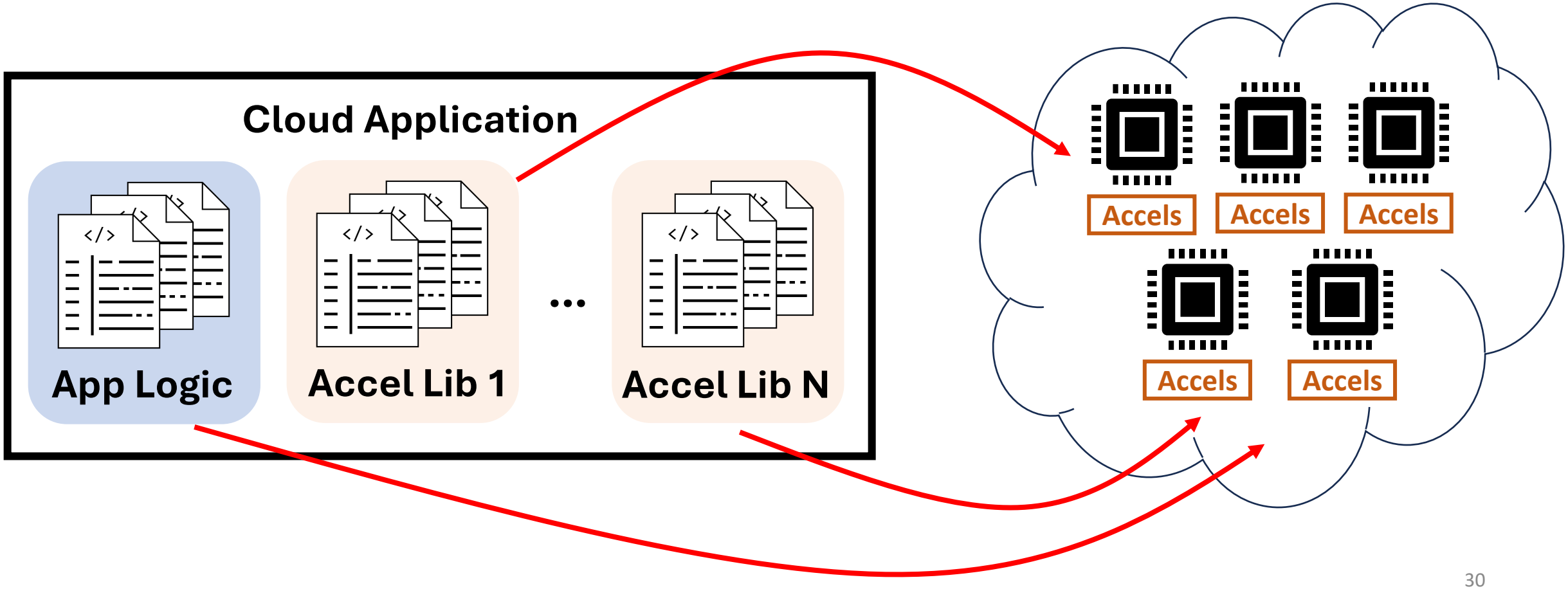
Limitation of SOTA Software Support

- Libraries still hard to use *efficiently*



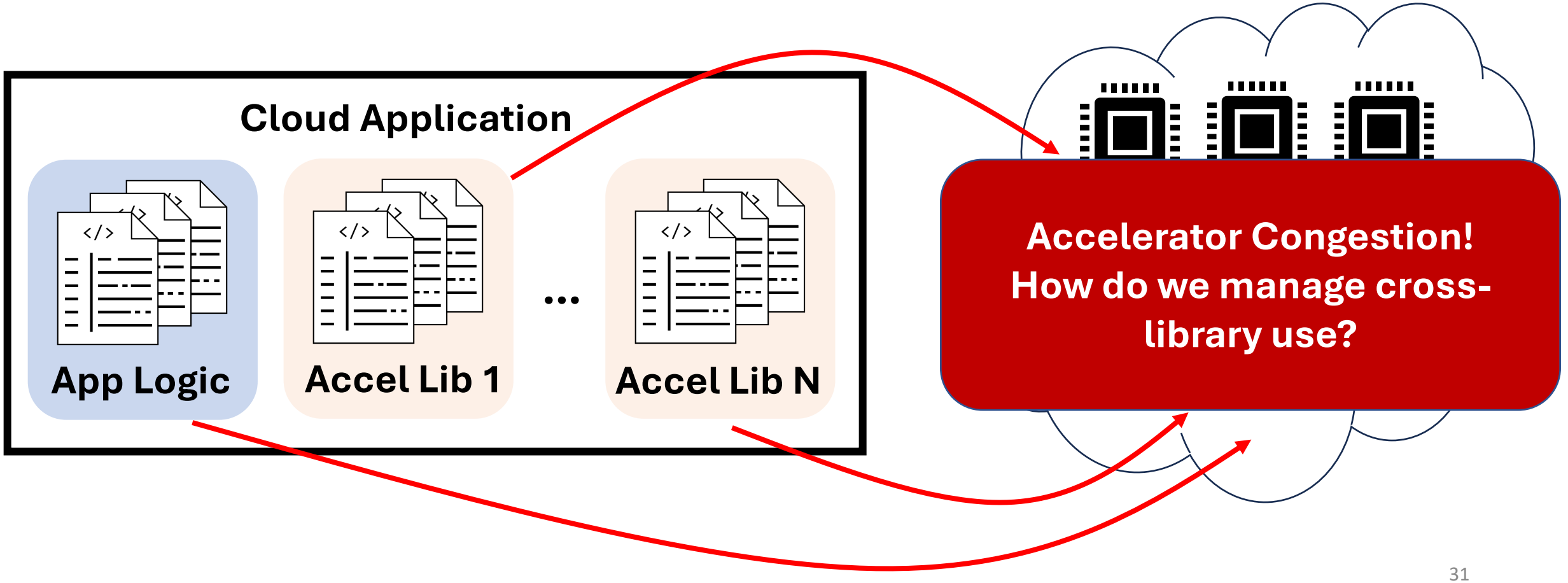
Limitation of SOTA Software Support

- Libraries still hard to use *efficiently*



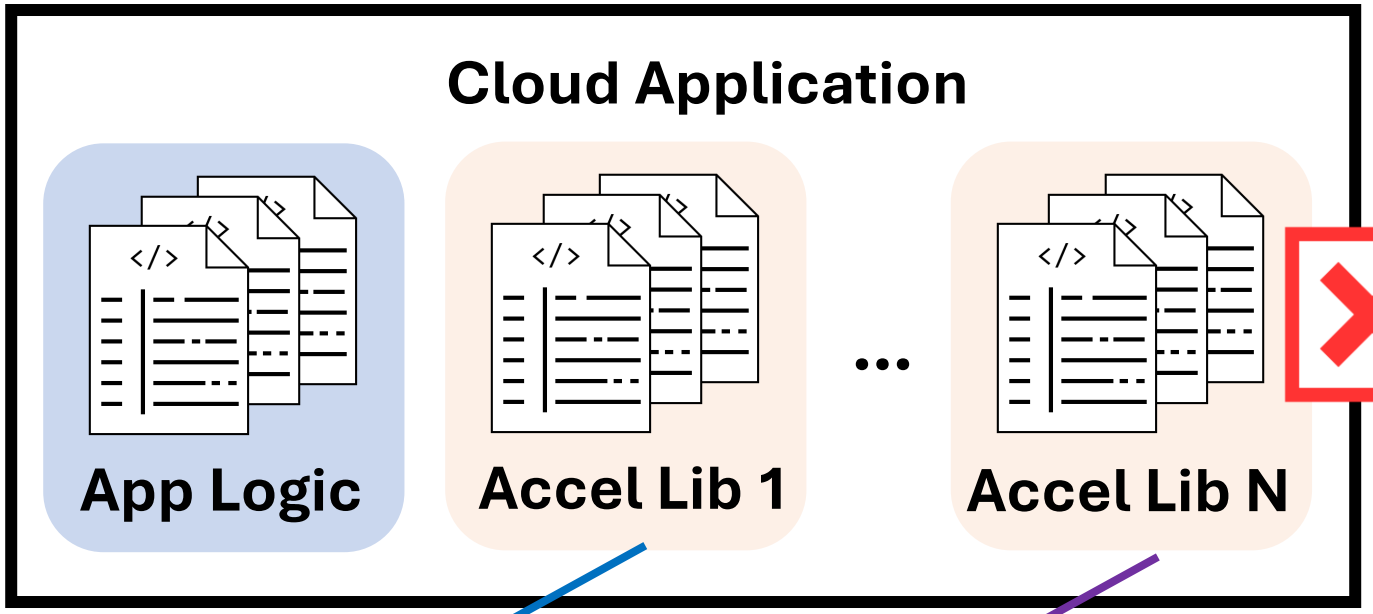
Limitation of SOTA Software Support

- Libraries still hard to use *efficiently*



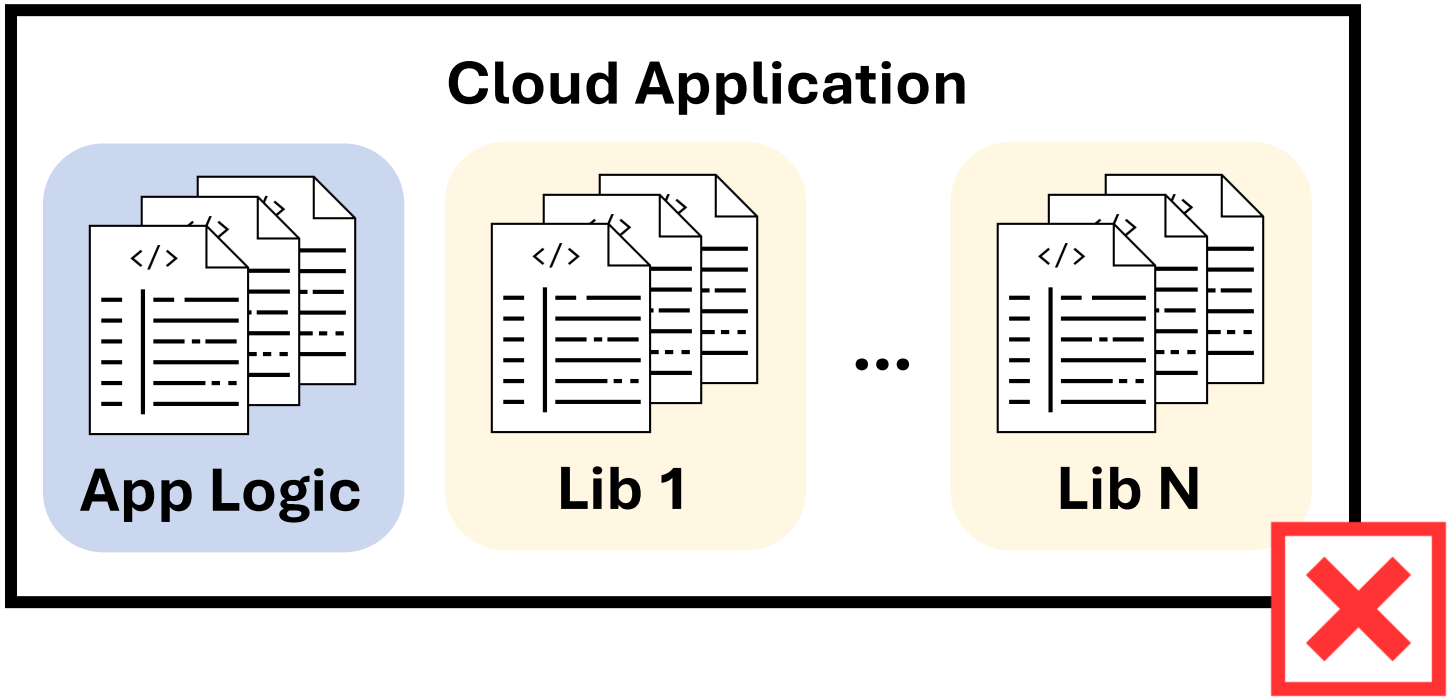
Limitation of SOTA Software Support

- Libraries still hard to use *efficiently*
- Compatibility



Unlocking Accelerators with AI

- Enables development *at scale*



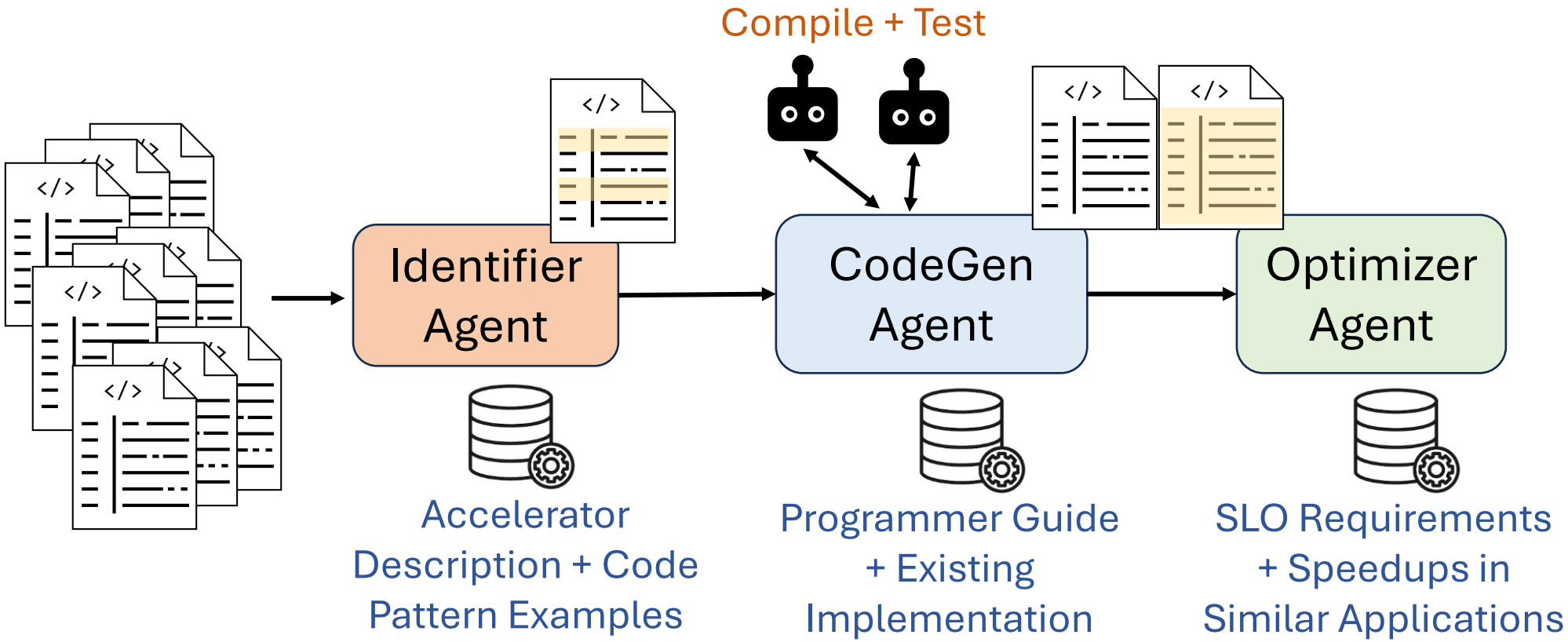
Unlocking Accelerators with AI

- Enables development *at scale*
- Enables *evolvable* accelerator support

Unlocking Accelerators with AI

- Enables development *at scale*
- Enables *evolvable* accelerator support
- Better addresses increase in complexity
  - Multiple Interfaces
  - Large Search Space of Runtime “Knobs”
  - Cross-library management and requirements

AccelAgent: Automatic Acceleration Framework



Conclusion

In summary, we:

- Identified a lack of utilization of on-chip accelerators in datacenters.
- Enumerated various challenges in efficient use of accelerators.
- Proposed an agentic framework for automated acceleration.

Hunting for Offload: Automated Discovery of Acceleratable Code in Datacenters

Joshua Kim[1], Chaojie Zhang[2], Íñigo Goiri[2], Christopher J. Rossbach[1,2], Jovan Stojkovic[1]

[1]The University of Texas at Austin, [2]Microsoft