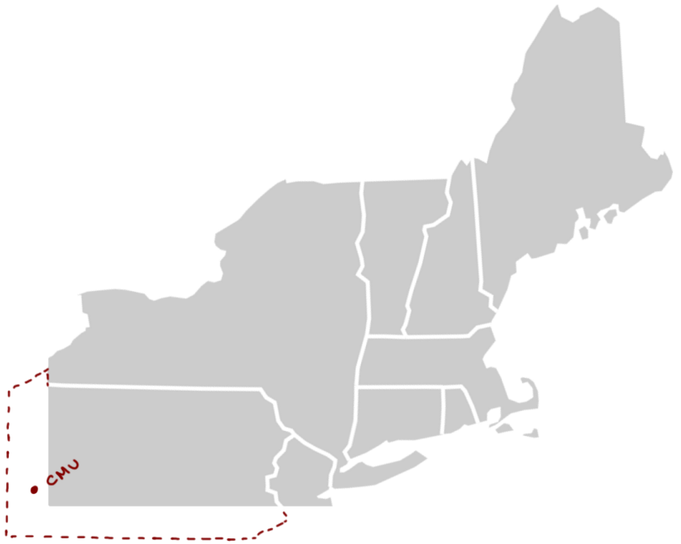
Embedding Database Logic in the Operating System Is Finally a Good Idea

Matthew Butrovich, Andrew Pavlo

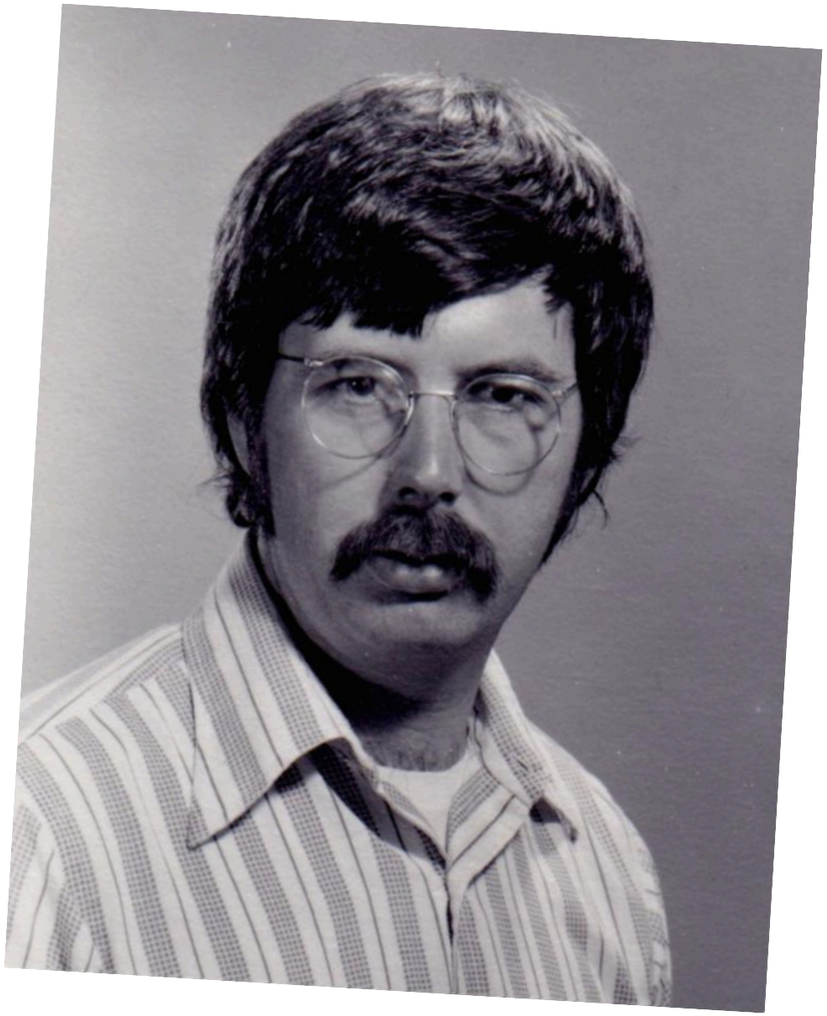


CARNEGIE MELLON DATABASE GROUP

NEDB Day 2023

We've Been Fighting the OS for Decades

"The bottom line is that operating system services in many existing systems are either too slow or inappropriate."

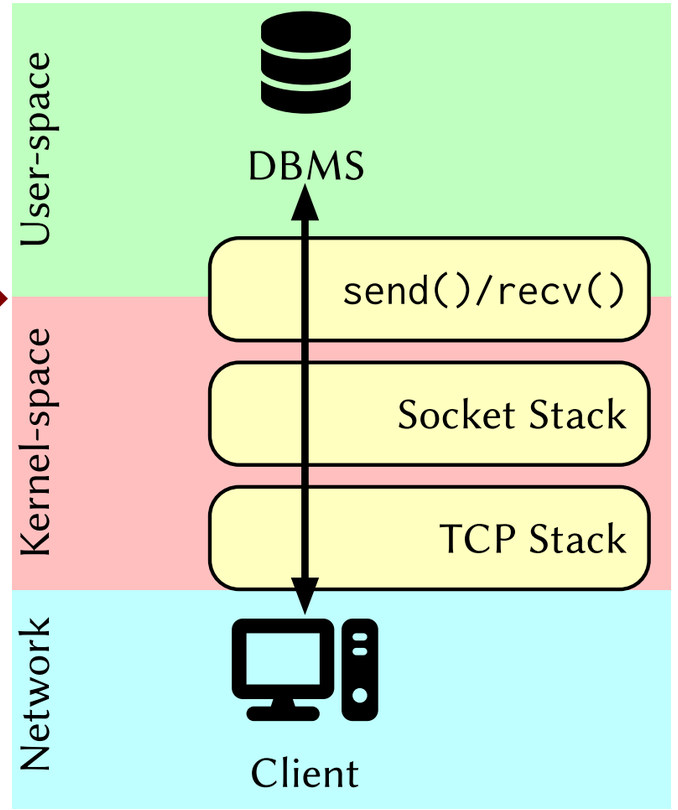


Michael Stonebraker. Operating System Support for Database Management. *Commun. ACM*. 1981.

Where Is the I/O Bottleneck?

- I/O devices (network, disk) are faster.
- Operating syster
- Max throughput: 42Gbps per CPU core.

>50% of CPU cycles on memcpy().

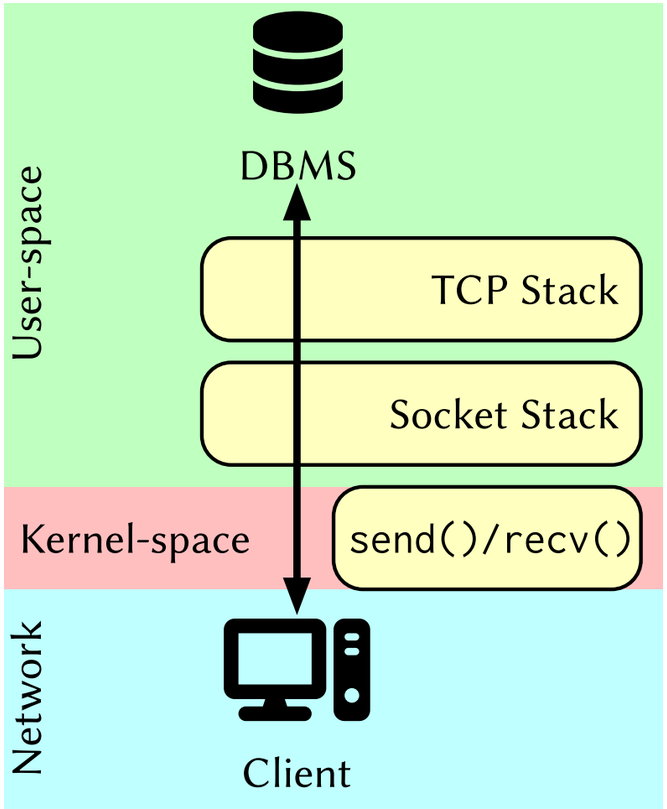


User-space DBMS

Qizhe Cai et al. Understanding host network stack overheads. *SIGCOMM*. 2021.

Kernel-Bypass

- Reimplement protocols in user-space.
- Difficult to debug, deploy, and maintain.
- Difficult to optimize.



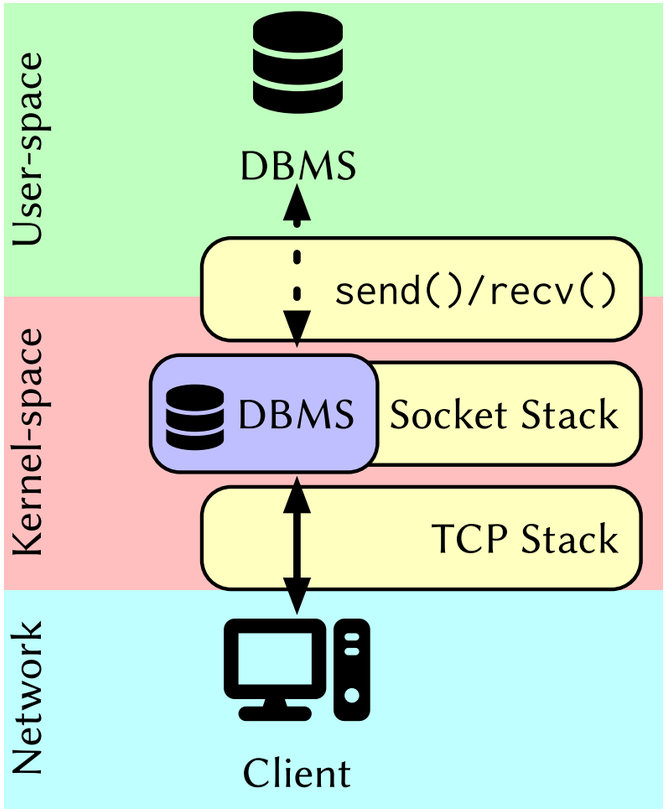
Kernel-bypass DBMS

William Tu et al. revisiting the openvSwitch dataplane ten years later. *SIGCOMM*. 2021.

https://github.com/xrp-project/BPF-KV/issues/3

User-Bypass

- Don’t pull DBMS data to user-space, push DBMS logic to kernel-space.
- Avoid copying buffers, scheduling user threads, and system call overhead.



User-bypass DBMS

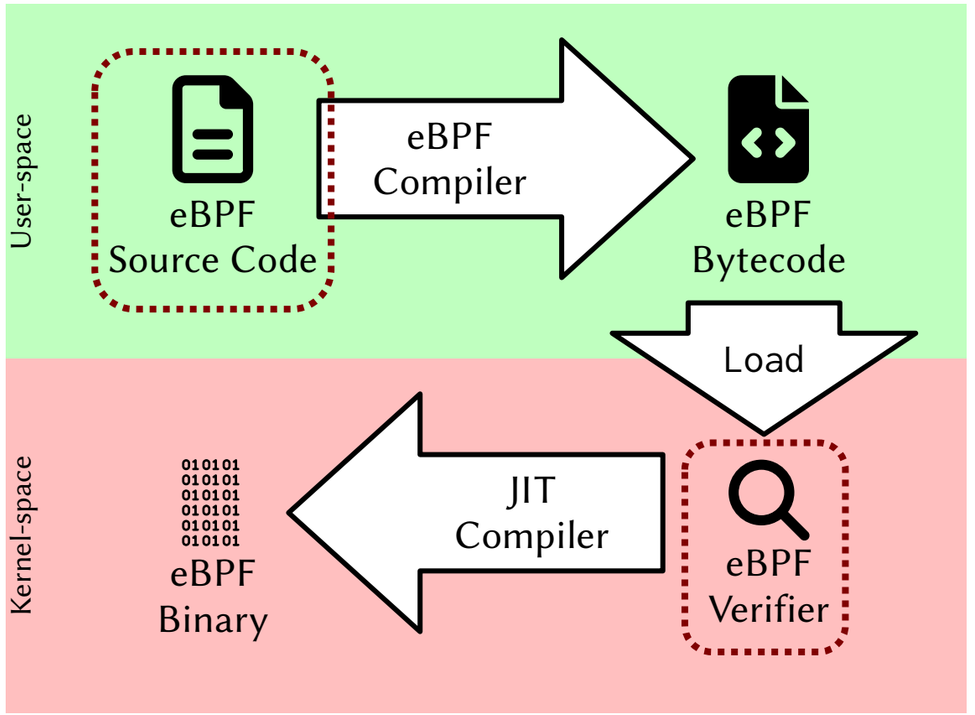
Brian N. Bershad et al. Extensibility, Safety and Performance in the SPIN Operating System. *SOSP*. 1995.

Felix Martin Schuhknecht et al. RUMA has it: Rewired User-space Memory Access is Possible! *VLDB*. 2016.

Margo I. Seltzer et al. Dealing with Disaster: Surviving Misbehaved Kernel Extensions. *OSDI*. 1996.

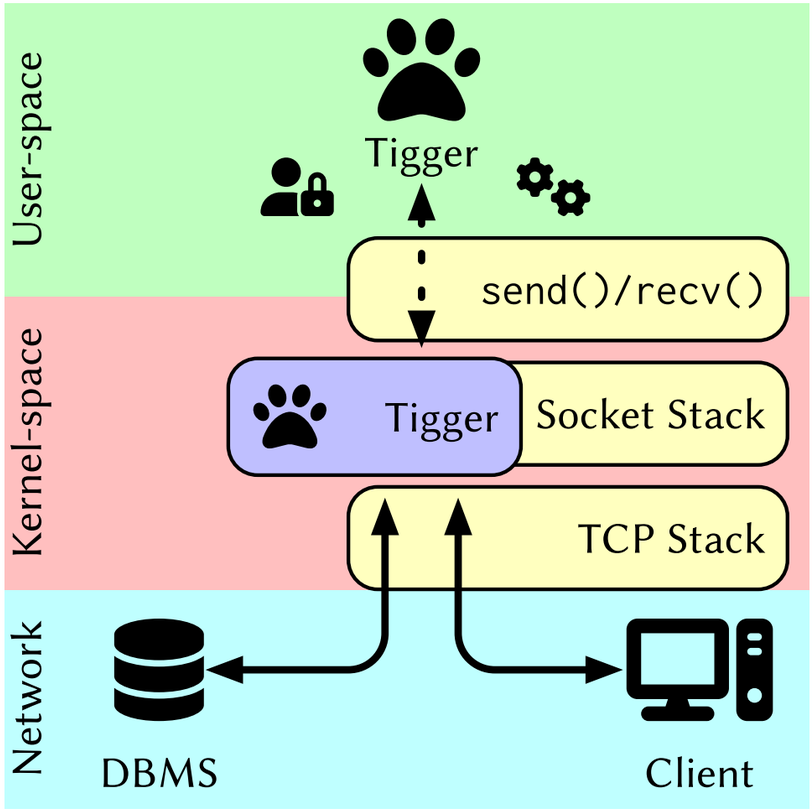
extended Berkeley Packet Filter

- Safe, event-driven programs in kernel-space
- Write in C and compile to eBPF
- Verifier constraints:
  - # instructions, boundedness, memory safety, limited API



DBMS Proxy with User-Bypass

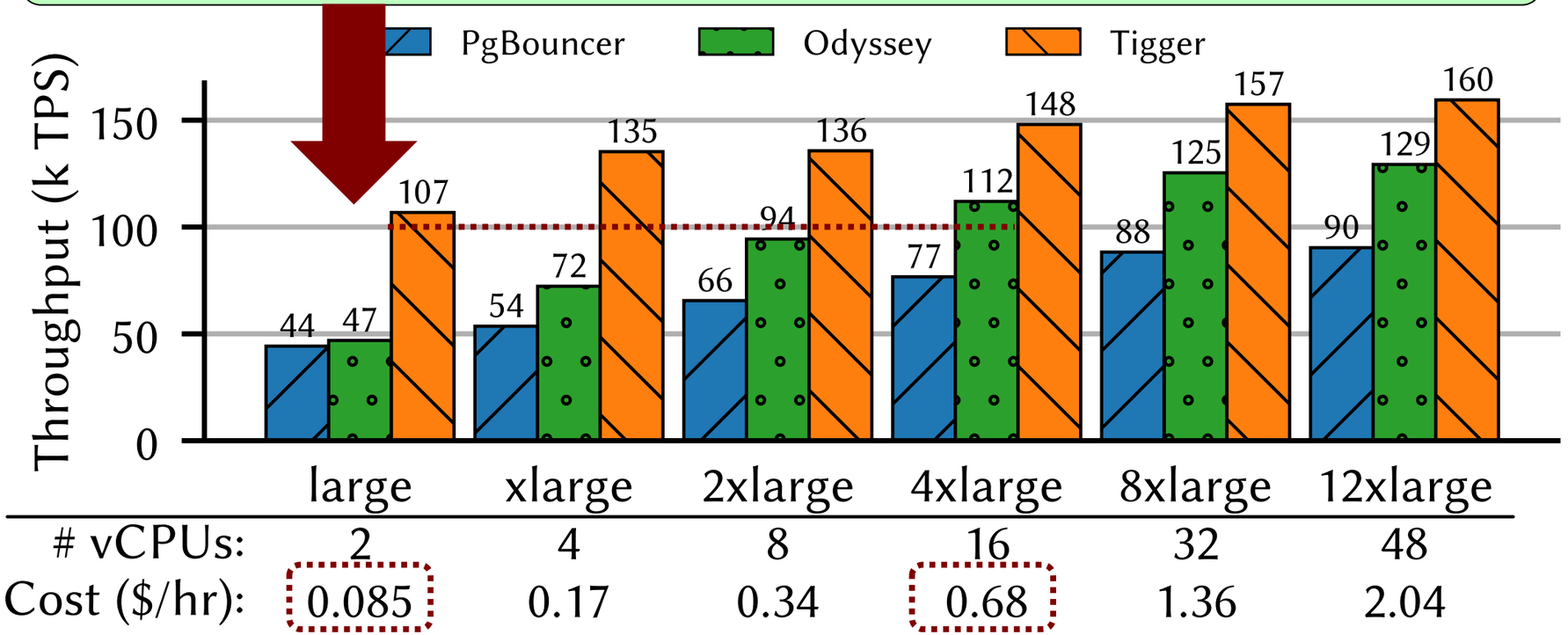
- User-bypass logic:
  - Transaction-aware pooling
  - Workload replication
- User-space component:
  - Authentication
  - Settings



User-Bypass DBMS Proxy

User-Bypass Performance

>2x throughput under severe CPU constraint

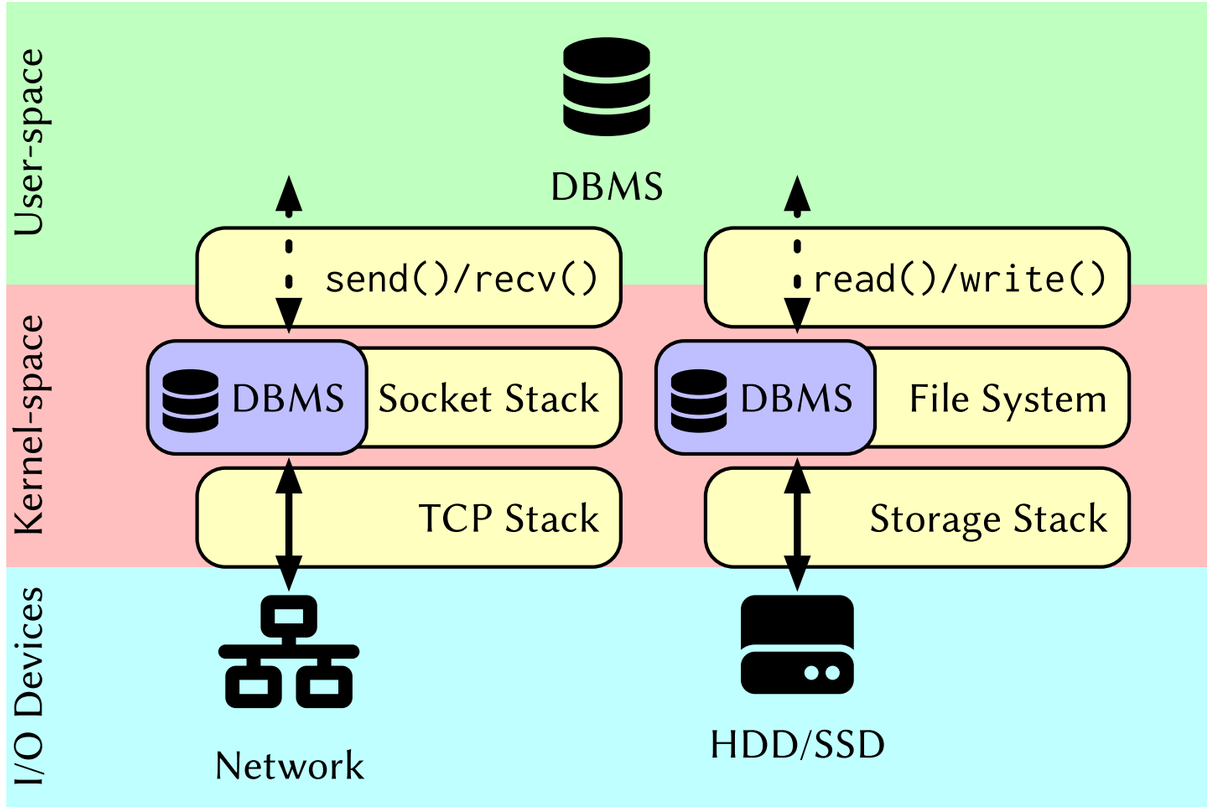


8x cost for Odyssey to match Tigger's performance

How to apply user-bypass to database systems?

Ephemeral I/Os

- Brought to user-space, processed, and discarded
- Storage
  - Index and MVCC traversal
  - Page filtering, garbage collection
- Networking:
  - Shuffle nodes



User-bypass DBMS

I will be on the job market later this year.

My wife wants to live in New England.

https://mattbutrovi.ch