香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Square: Towards Sharing Internet-Scale GPU Clouds Fairly and Efficiently

Ningxin Su, Freeman Cheng, Baochun Li

Ningxin Su, Assistant Professor
Information Hub, IoT Thrust
HKUST (GZ)

IWQoS 2026, Istanbul

香港科技大學（廣州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

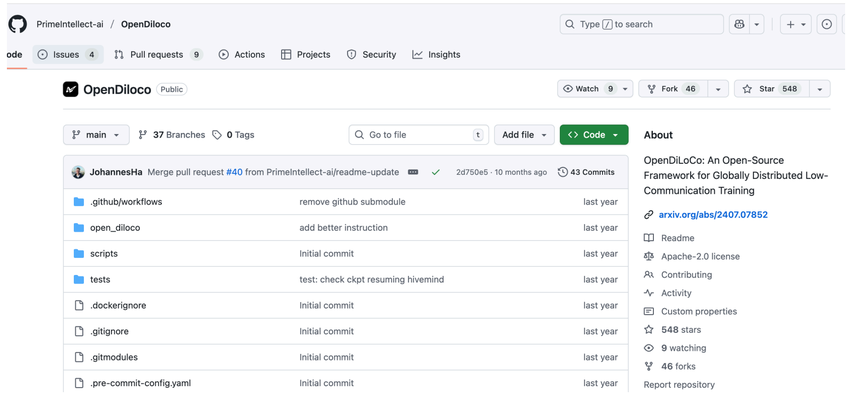
Google DeepMind

- DiLoCo
- a variant of FedAvg

https://arxiv.org/abs/2311.08105

PrimeIntellect-ai

- OpenDiLoCo
- Open-Source

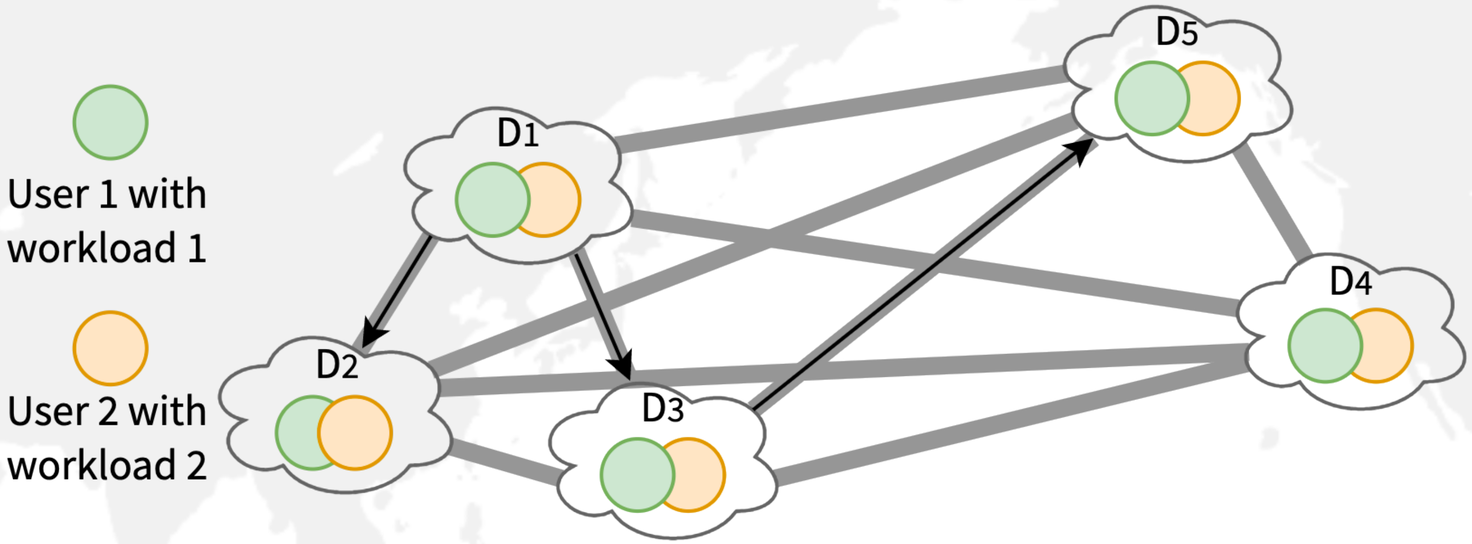


https://github.com/PrimeIntellect-ai/OpenDiloco

Gongji Cloud

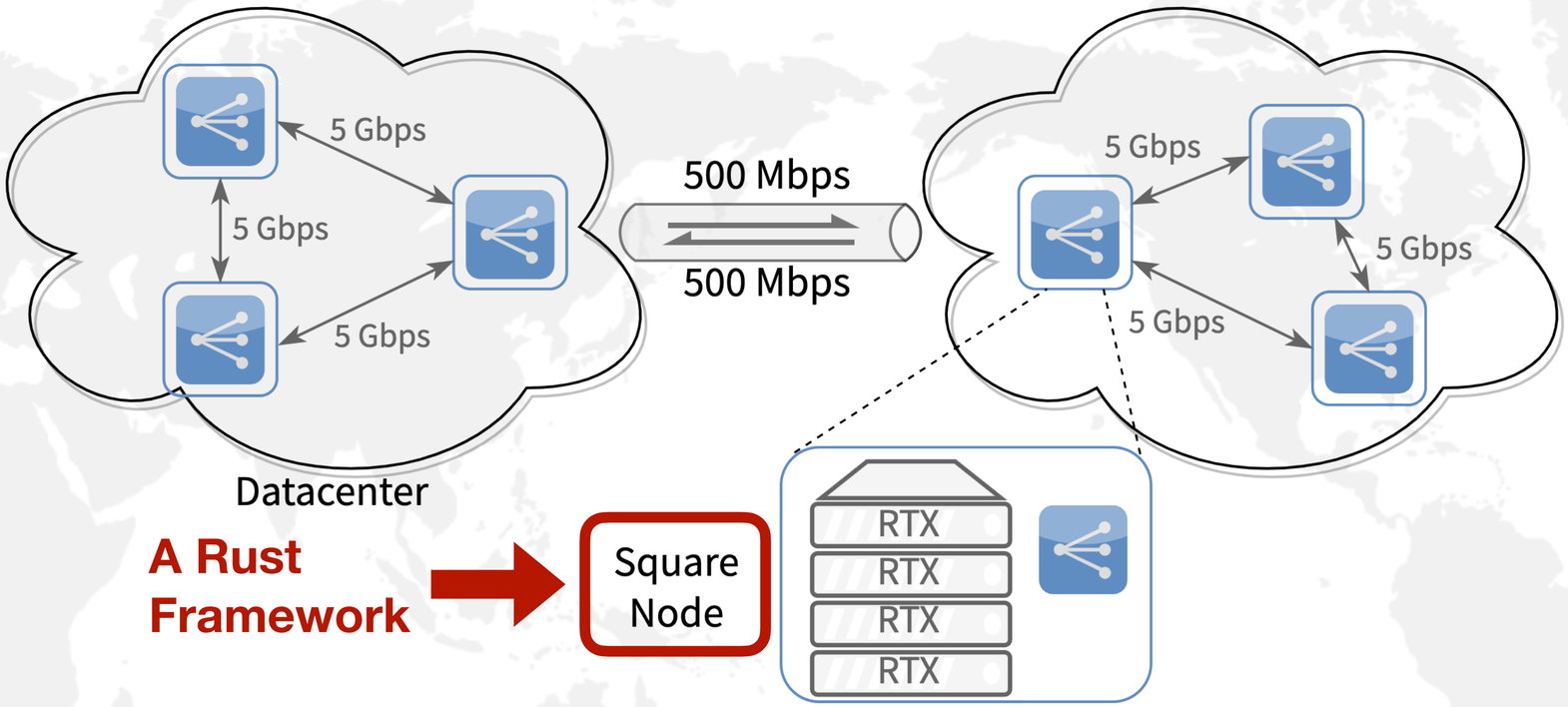
https://suanli.cn/

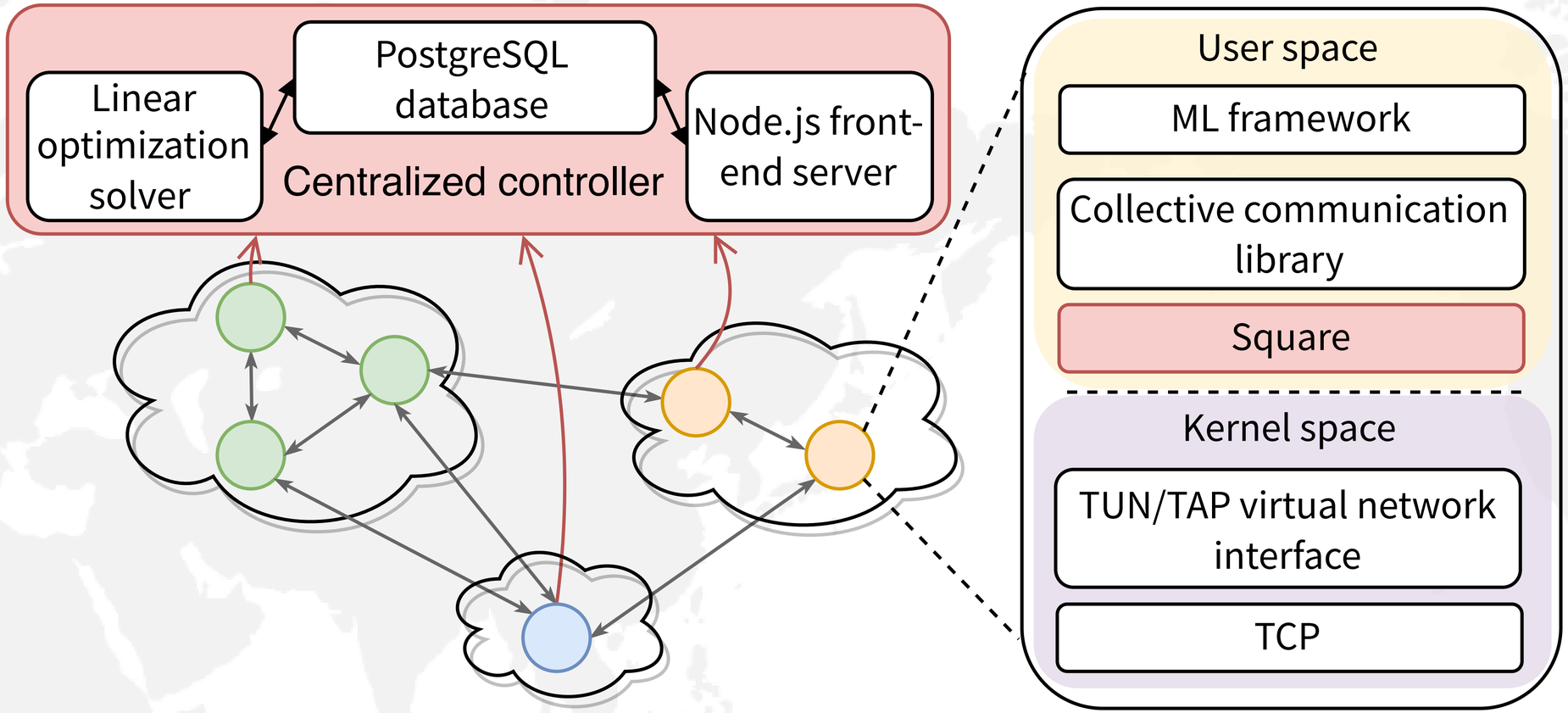
香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

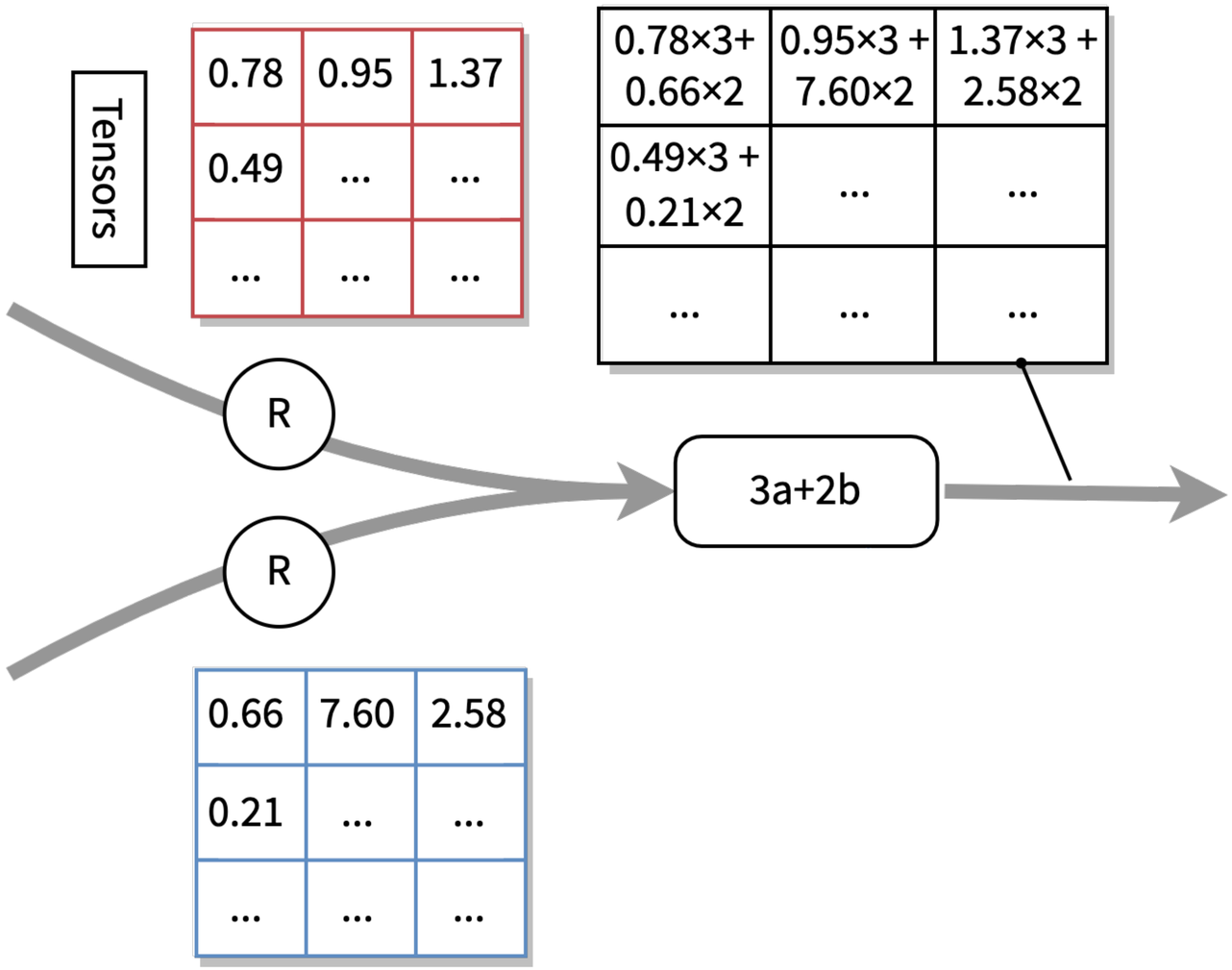


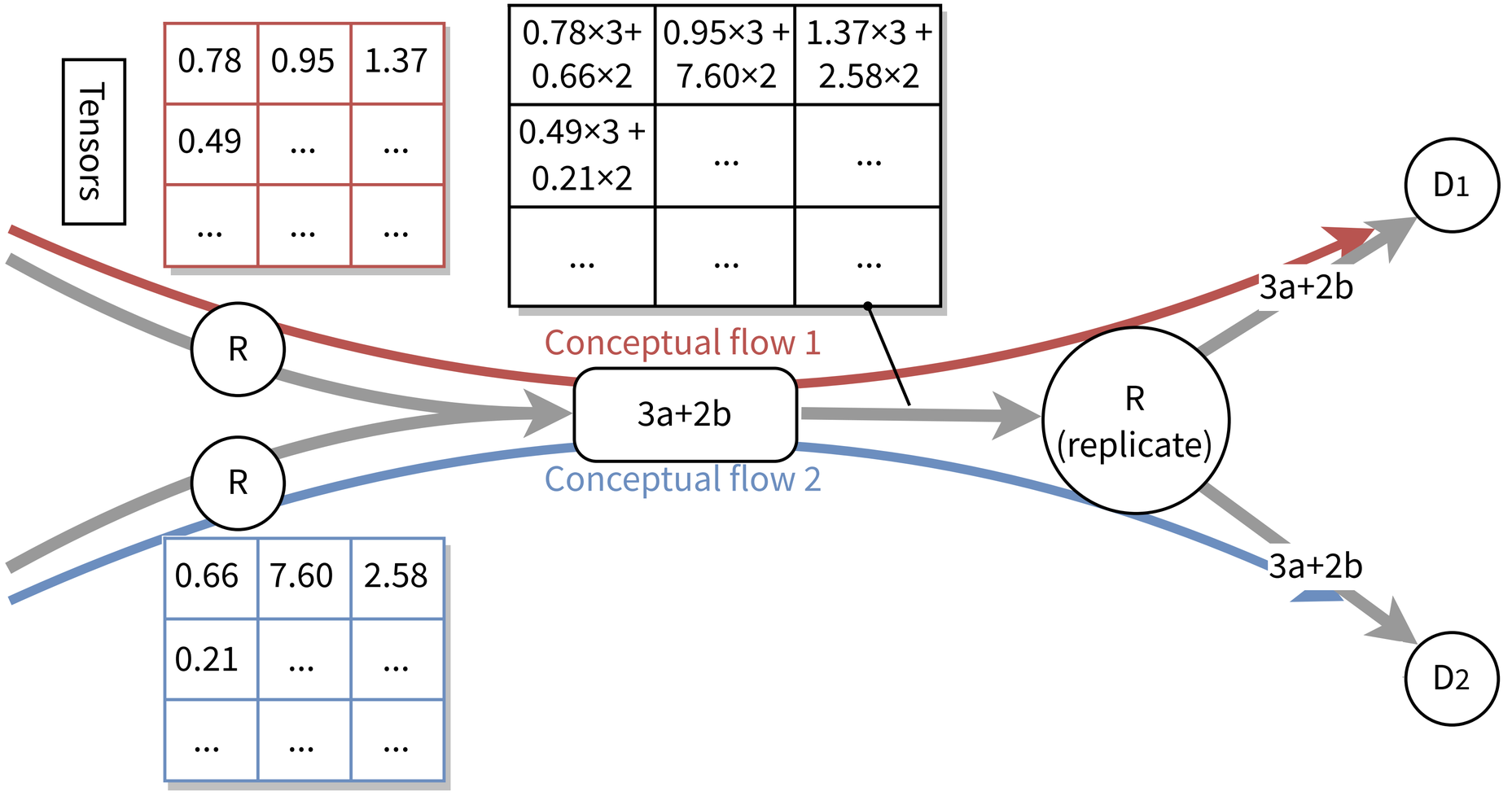
When many users train across many cloud regions, how should scarce WAN bandwidth be shared fairly and efficiently?

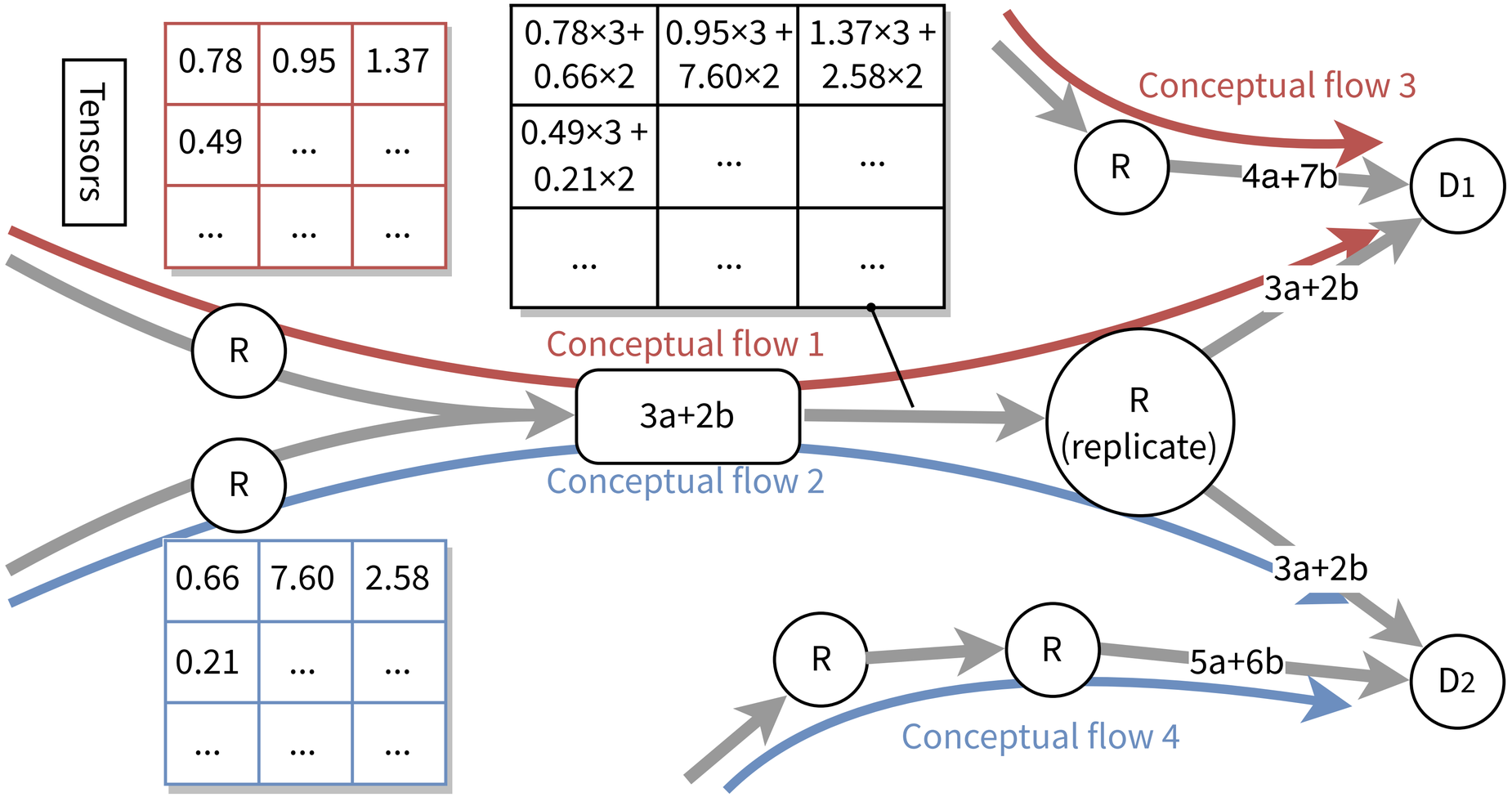
香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

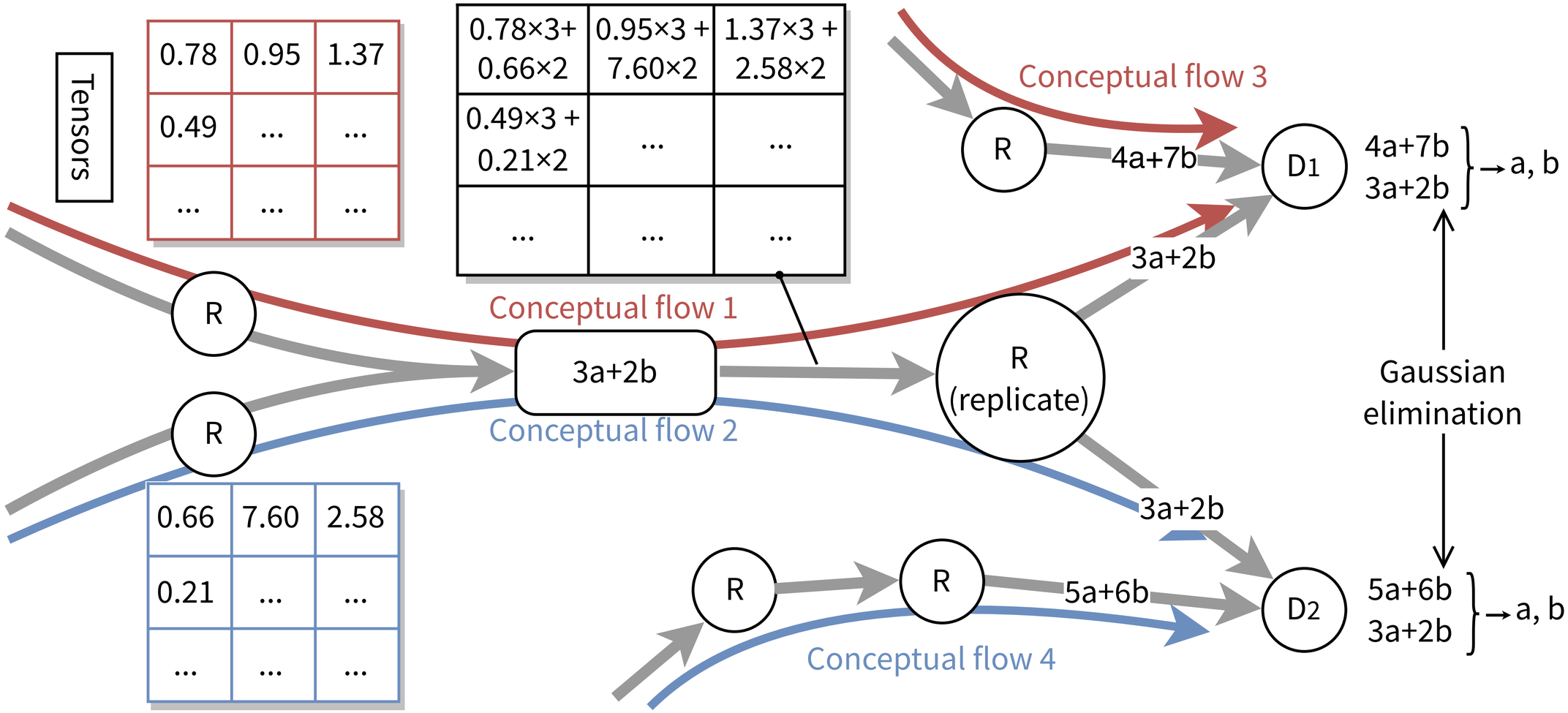



Linear optimization solver
PostgreSQL database
Node.js front-end server
Centralized controller
User space
ML framework
Collective communication library
Square
Kernel space
TUN/TAP virtual network interface
TCP

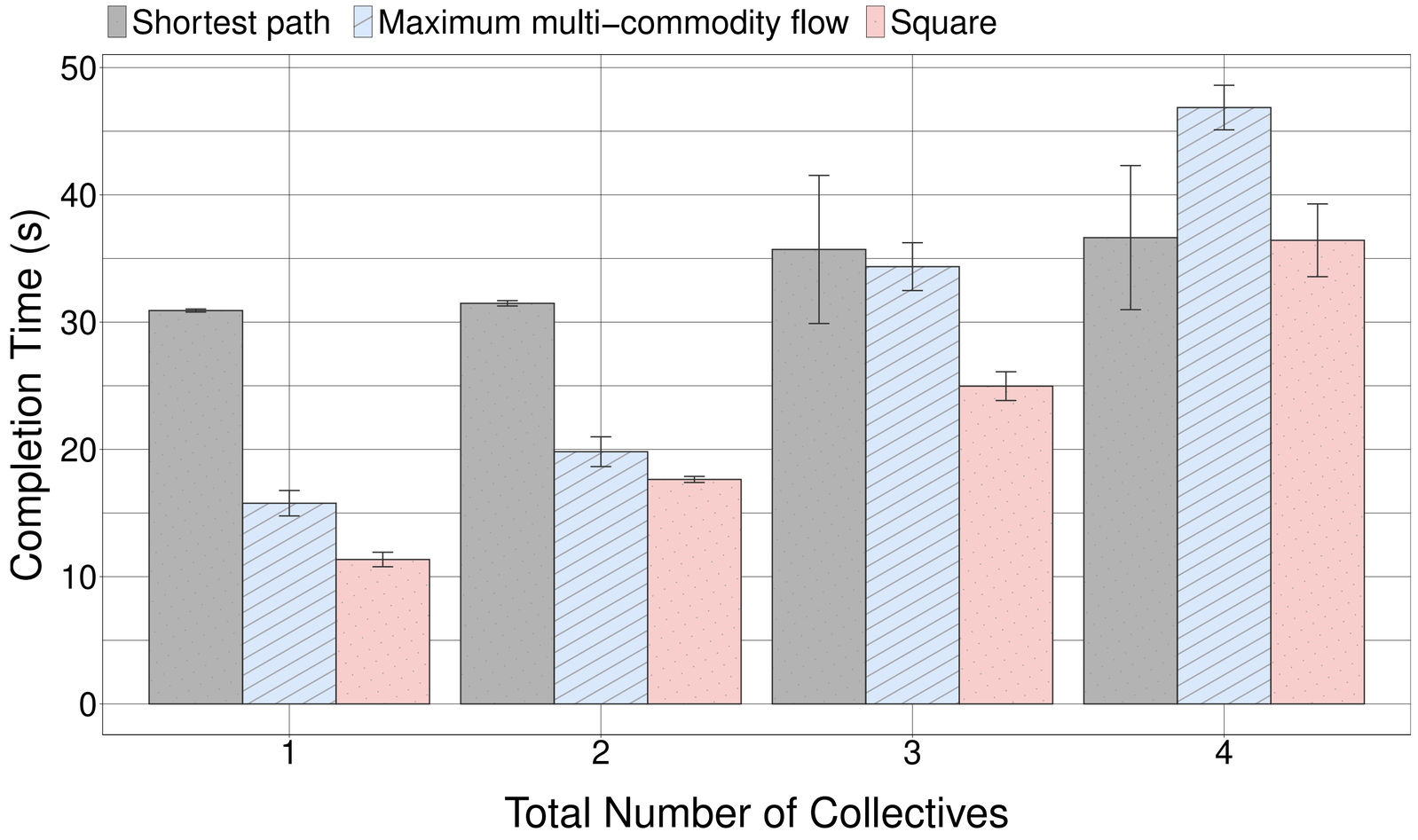

Tensors
0.78 0.95 1.37
0.49 ... ...
... ...
R
R
3a+2b
0.66 7.60 2.58
0.21 ... ...
... ...
0.78×3+ 0.66×2
0.95×3 + 7.60×2
1.37×3 + 2.58×2
0.49×3 + 0.21×2
...
...
...
...
...


Tensors
0.78 0.95 1.37
0.49 ... ...
... ...
R
R
0.66 7.60 2.58
0.21 ... ...
... ...
0.78×3+ 0.66×2
0.95×3 + 7.60×2
1.37×3 + 2.58×2
0.49×3 + 0.21×2
...
Conceptual flow 1
Conceptual flow 2
3a+2b
R (replicate)
3a+2b
3a+2b
D1
D2


Tensors
0.78 0.95 1.37
0.49 ... ...
... ...
0.66 7.60 2.58
0.21 ... ...
... ...
0.78×3+ 0.66×2
0.95×3 + 7.60×2
1.37×3 + 2.58×2
0.49×3 + 0.21×2
...
R
R
3a+2b
Conceptual flow 1
Conceptual flow 2
Conceptual flow 3
R
4a+7b
D1
3a+2b
R (replicate)
3a+2b
R
R
5a+6b
D2
Conceptual flow 4


Tensors
0.78 0.95 1.37
0.49 ... ...
... ...
0.66 7.60 2.58
0.21 ... ...
... ...
R
R
3a+2b
0.78×3+ 0.66×2
0.95×3 + 7.60×2
1.37×3 + 2.58×2
0.49×3 + 0.21×2
Conceptual flow 1
Conceptual flow 2
Conceptual flow 3
Conceptual flow 4
R
4a+7b
D1
4a+7b
3a+2b
a, b
3a+2b
R (replicate)
3a+2b
Gaussian elimination
R
R
5a+6b
D2
5a+6b
3a+2b
a, b

香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)



Experimental results over a real Digital Ocean inter-datacenter overlay.

香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Thank you!

ningxinsu@hkust-gz.edu.cn

nebulis-lab.com

ningxinsu.github.io