Scaling Up Private Information Retrieval

David Wu

October 2026



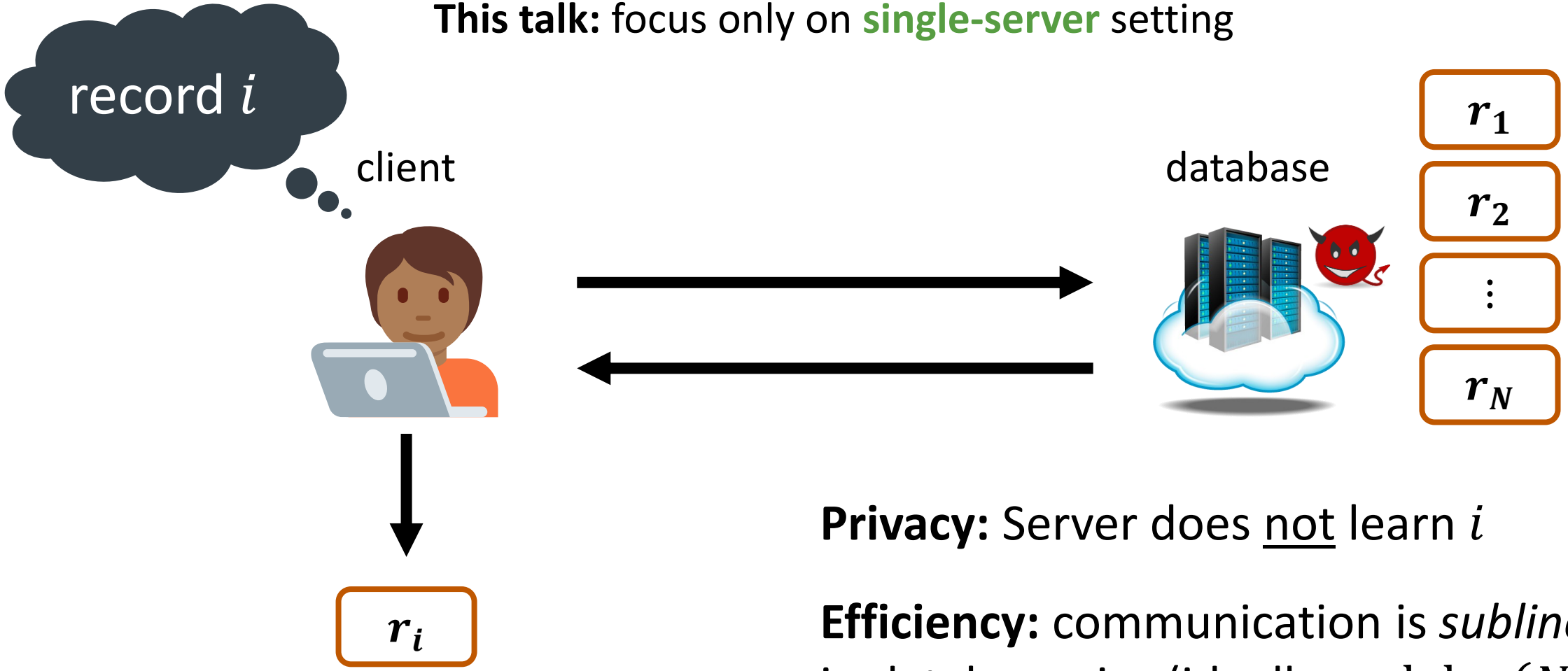


based on joint works with Samir Menon and Sidaarth Sabhnani

Private Information Retrieval (PIR)

[CGKS95]

This talk: focus only on **single-server** setting



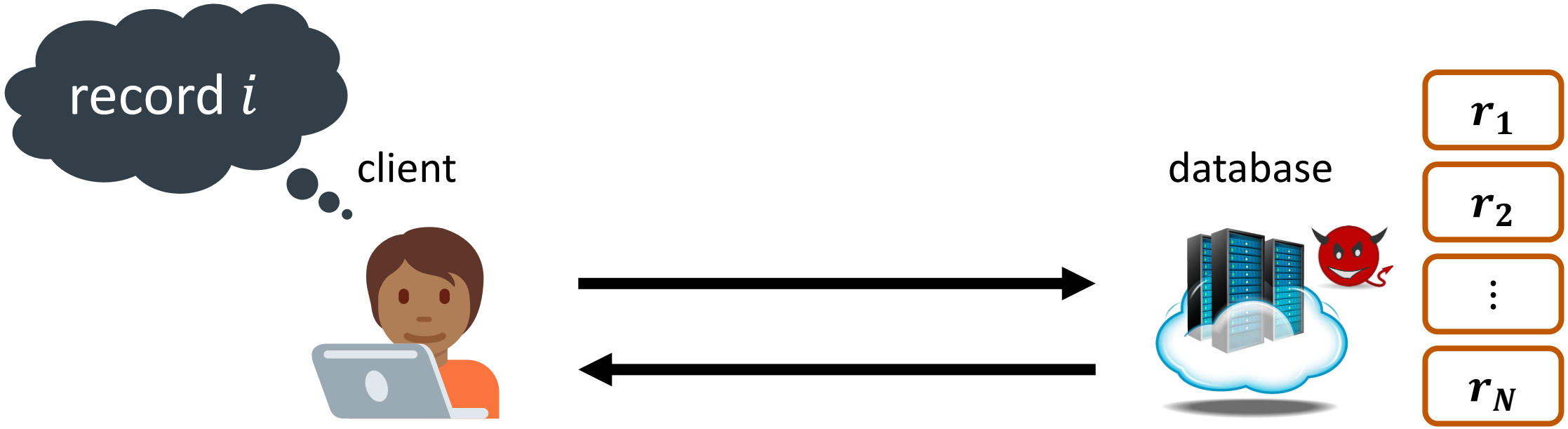
Privacy: Server does not learn i

Efficiency: communication is *sublinear* in database size (ideally: $\text{polylog}(N)$)

Correctness: Client learns r_i

Private Information Retrieval (PIR)

[CGKS95]

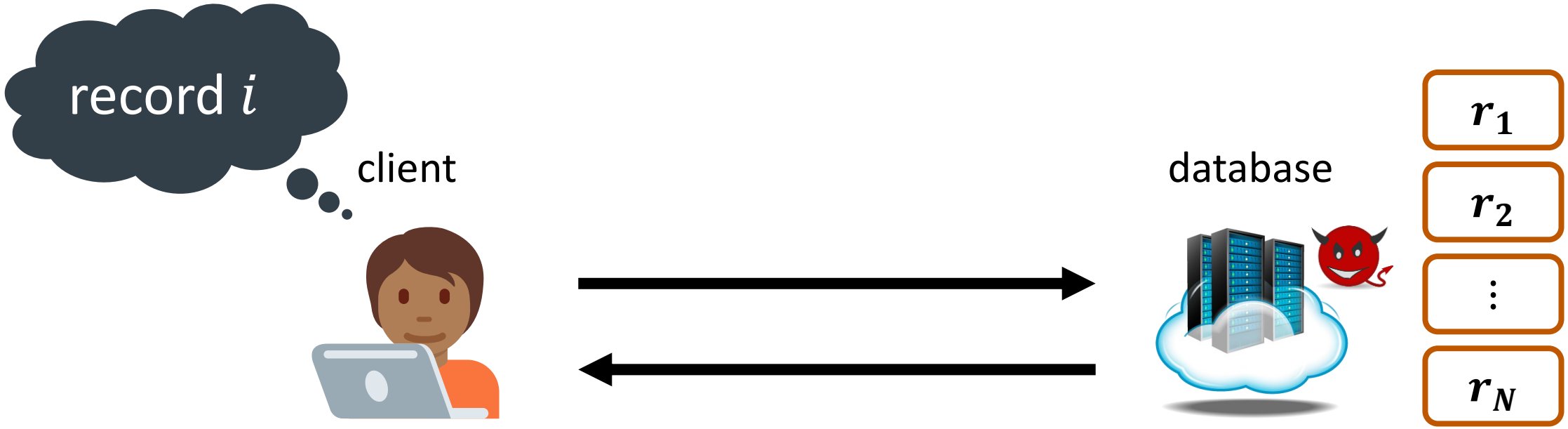


Basic building block in many privacy-preserving protocols

- Metadata-private messaging
- Certificate transparency auditing
- Private content delivery
- Contact discovery
- Private web search
- Private navigation
- Private contact tracing
- Private DNS
- Private blocklist lookup

Private Information Retrieval (PIR)

[CGKS95]

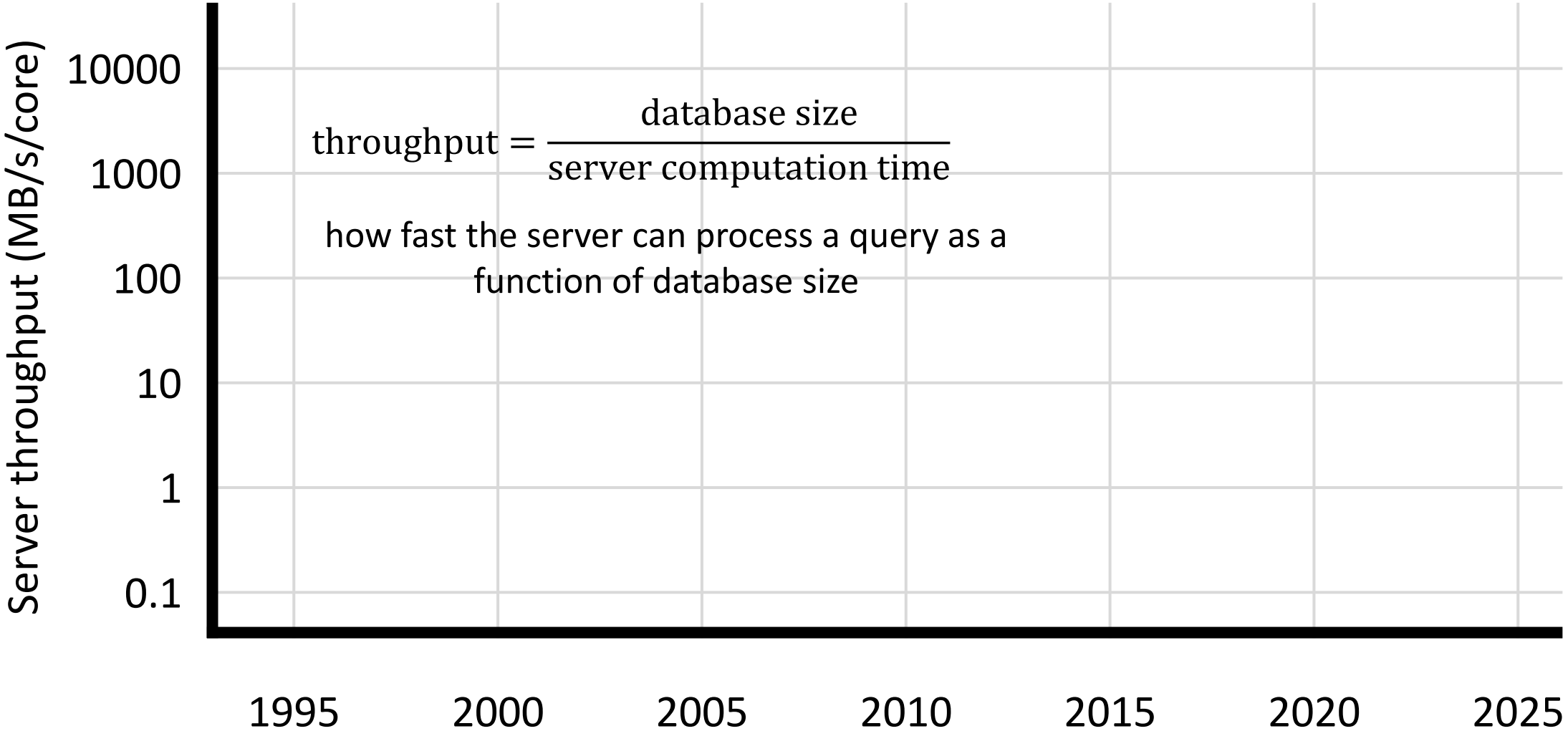


Basic building block in many privacy-preserving protocols

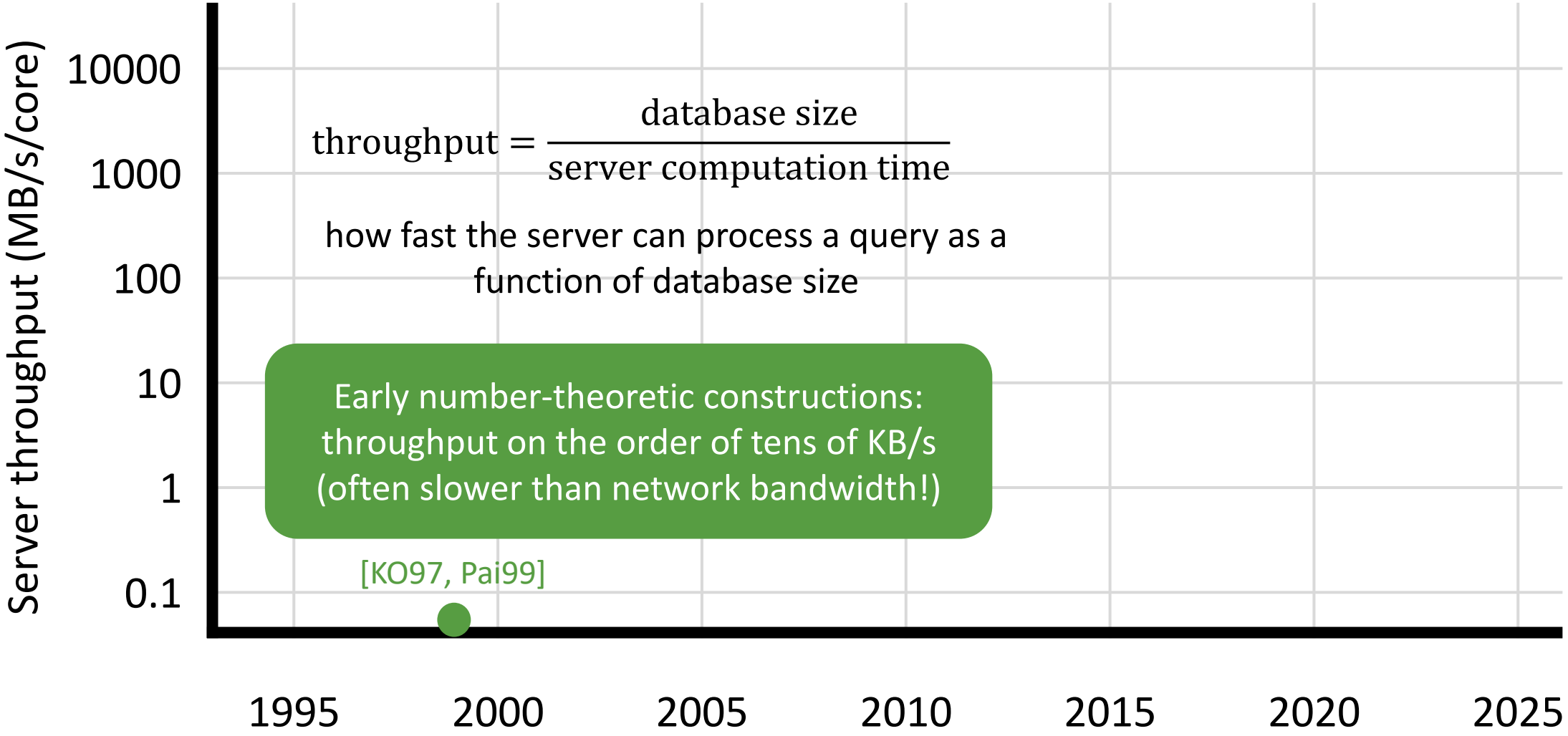
Recent deployments:

- Apple: private caller ID lookup, private image search
- Meta: advanced browsing protection

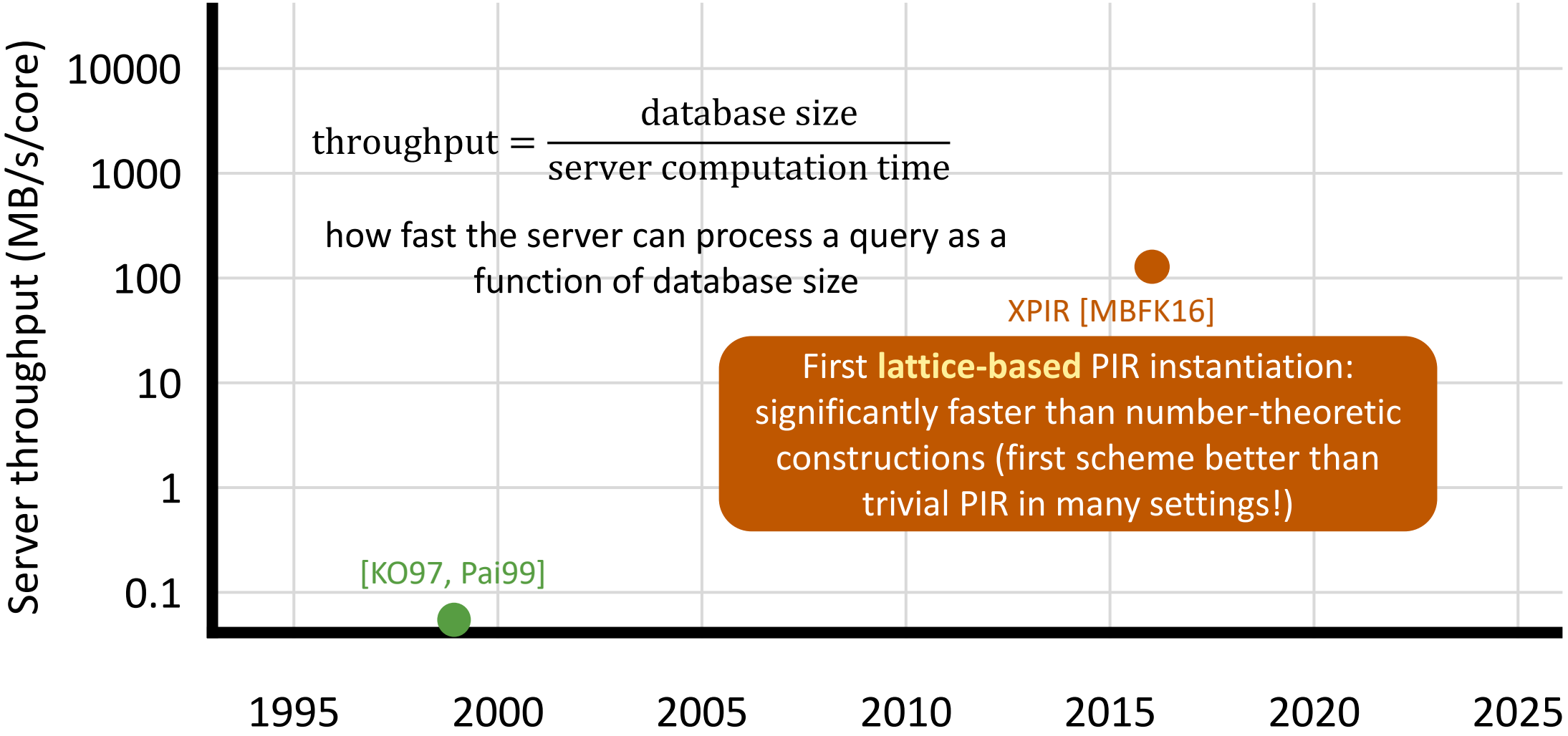
30 Years of PIR Research



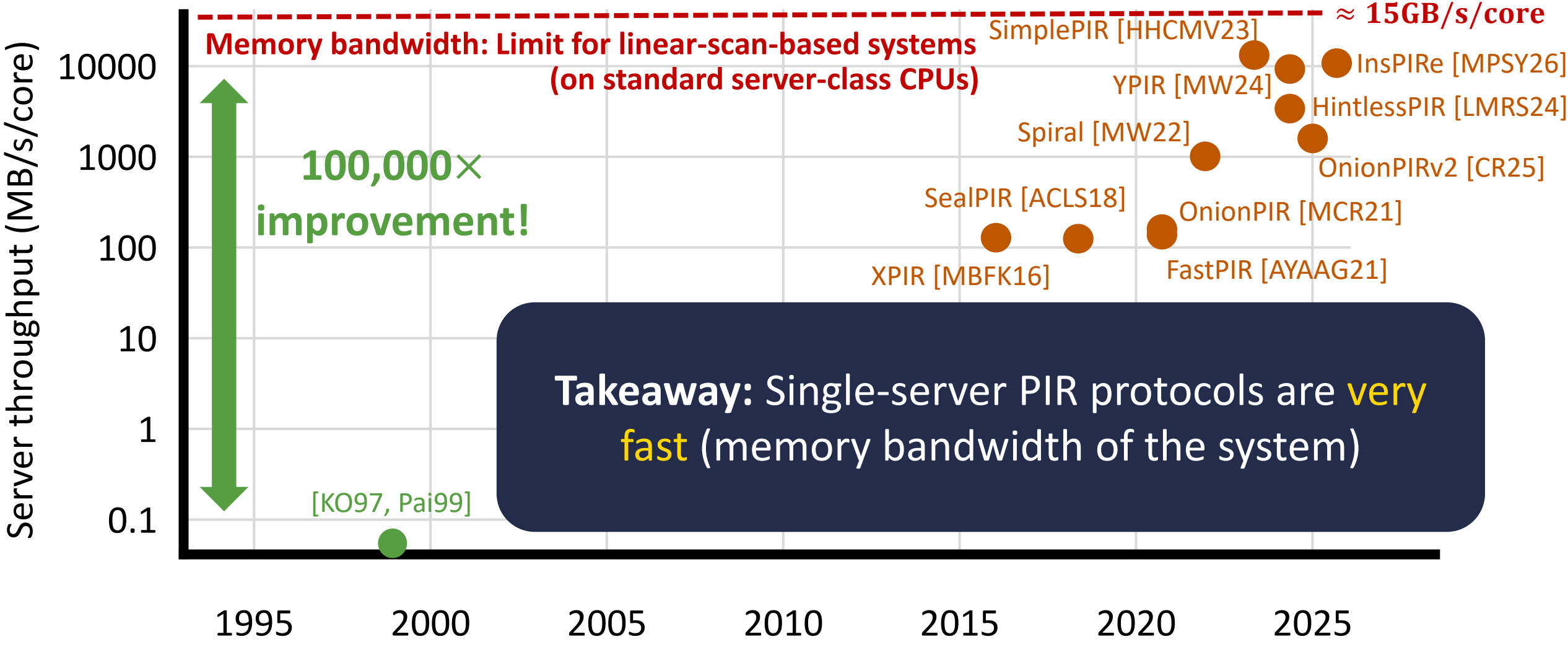
30 Years of PIR Research



30 Years of PIR Research



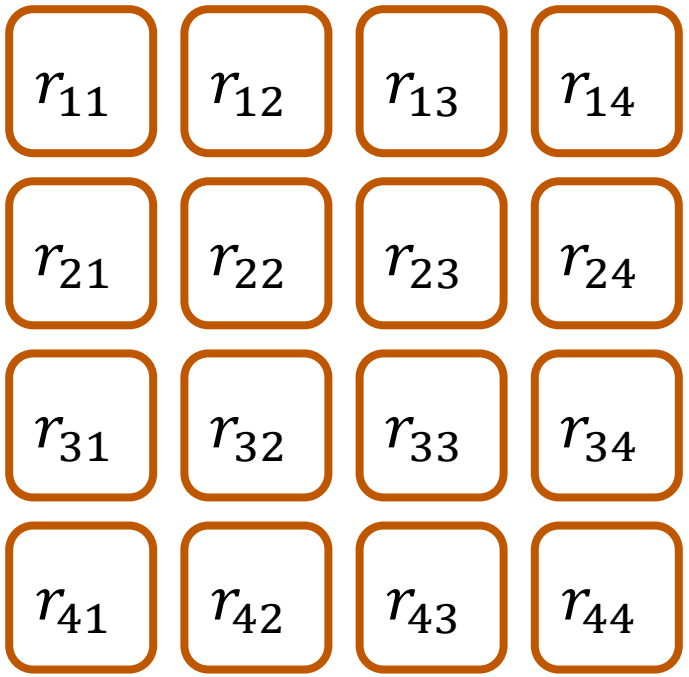
30 Years of PIR Research



PIR from Homomorphic Encryption

[KO97]

Starting point: a $\sqrt{N}$ construction (N = number of records)

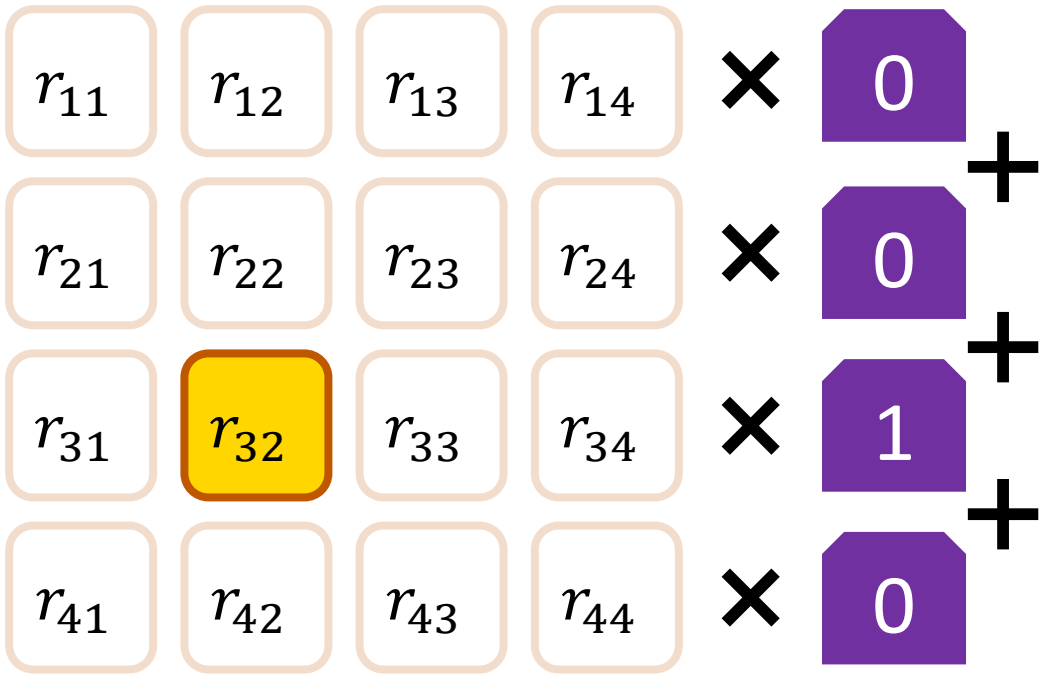


Arrange the database as a $\sqrt{N}$-by-$\sqrt{N}$ matrix

PIR from Homomorphic Encryption

[KO97]

Starting point: a $\sqrt{N}$ construction (N = number of records)



Encrypt a 0/1 vector indicating the row containing the desired record

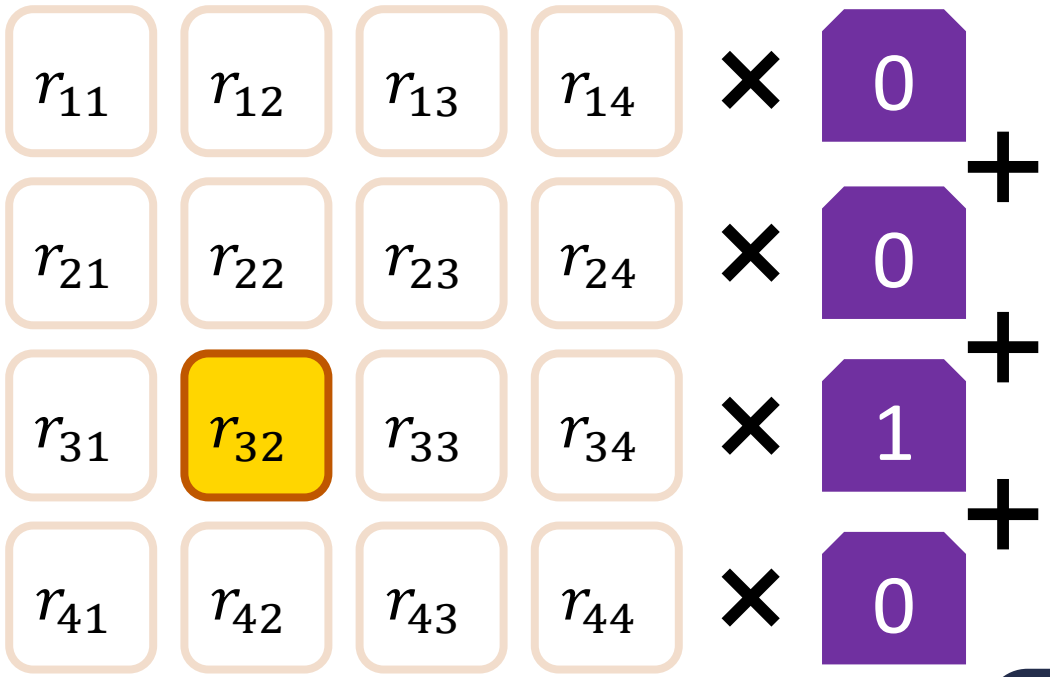
Arrange the database as a $\sqrt{N}$-by-$\sqrt{N}$ matrix

Homomorphically compute product between query vector and database matrix

PIR from Homomorphic Encryption

[KO97]

Starting point: a $\sqrt{N}$ construction (N = number of records)



Encrypt a 0/1 vector indicating the row containing the desired record



Arrange the database as a $\sqrt{N}$-by-$\sqrt{N}$ matrix

Database is in the clear, so *additive* homomorphism suffices

PIR from Homomorphic Encryption

[KO97]

Starting point: a $\sqrt{N}$ construction (N = number of records)

Client decrypts to learn records



Encrypt a 0/1 vector indicating the row containing the desired record

r_{31} r_{32} r_{33} r_{34}

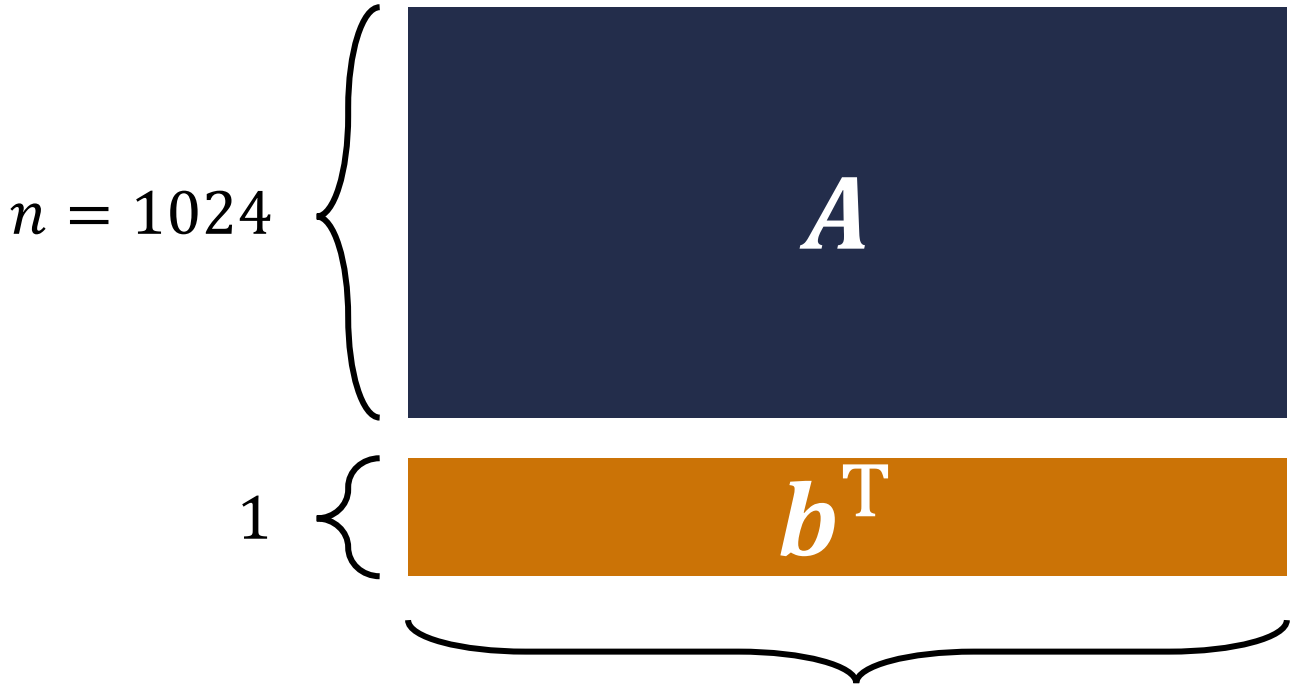
Response size: $O_\lambda(\sqrt{N})$

Homomorphically compute product between query vector and database matrix

SimplePIR: Lightweight PIR from LWE

[HHCMV23]

Building block: Regev's linearly homomorphic encryption from LWE [Reg05]



Elements of $\boldsymbol{A}$ are random 32-bit integers

$\boldsymbol{b}$ is a ciphertext (encrypts an indicator vector)

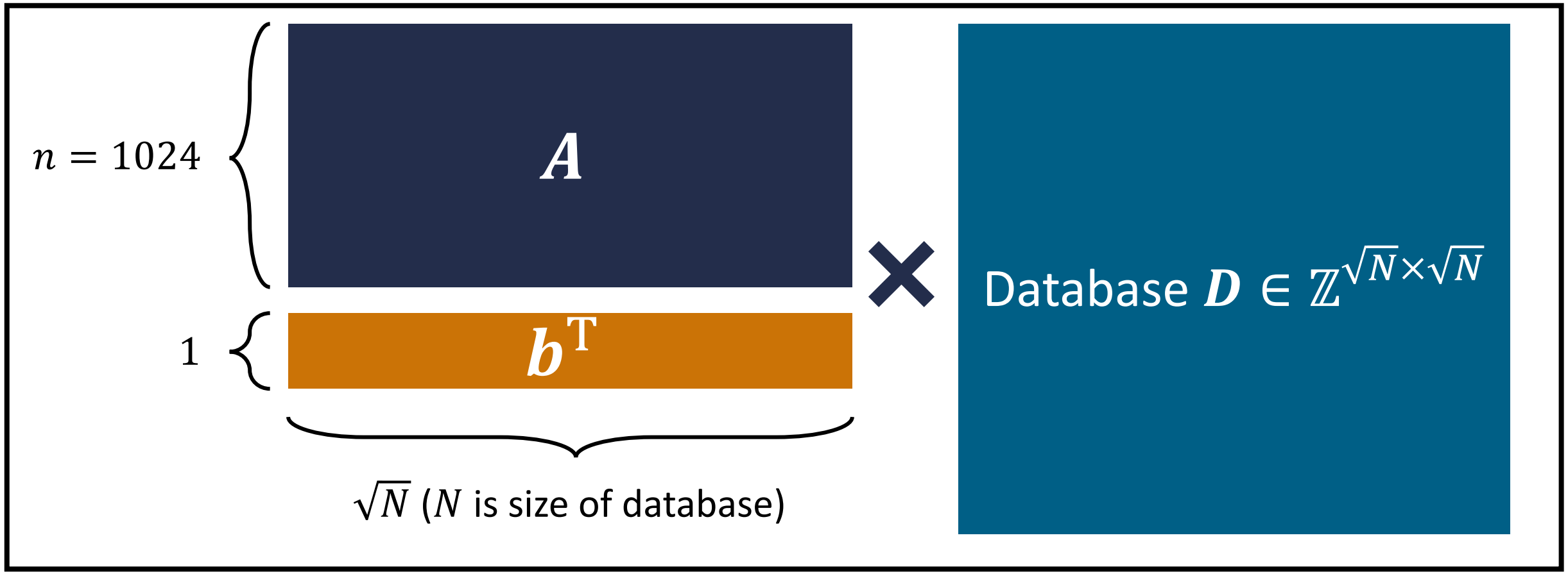
Query structure in SimplePIR

All components are elements of $\mathbb{Z}_q = \mathbb{Z}_{2^{32}}$ (i.e., 32-bit integers)

SimplePIR: Lightweight PIR from LWE

[HHCMV23]

Building block: Regev's linearly homomorphic encryption from LWE [Reg05]



All components are elements of $\mathbb{Z}_q = \mathbb{Z}_{2^{32}}$ (i.e., 32-bit integers)

SimplePIR: Lightweight PIR from LWE

[HHCMV23]

Key insight in SimplePIR: $\boldsymbol{A}$ is query-independent so move computation of $\boldsymbol{AD}$ offline

Given $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}$ and $\boldsymbol{AD}$, client can decrypt and learn the record



$\boldsymbol{AD}$

offline phase

online phase

$\boldsymbol{b}^{\mathrm{T}}$

$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}$



Database

$\boldsymbol{D} \in \mathbb{Z}_p^{\sqrt{N}\times\sqrt{N}}$

Query and response size:

$\sqrt{N}$ elements over $\mathbb{Z}_q$

Server computation:

$2N$ operations over $\mathbb{Z}_q$

2 operations per database entry

SimplePIR: Lightweight PIR from LWE

[HHCMV23]

Key insight in SimplePIR: $\boldsymbol{A}$ is query-independent so move computation of $\boldsymbol{AD}$ offline

Limitation: hint is **large** ($n\times$ larger than response) and needs to be updated whenever database changes



$\boldsymbol{AD}$

offline phase

online phase

$\boldsymbol{b}^{\mathrm{T}}$

$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}$



Database $\boldsymbol{D} \in \mathbb{Z}_p^{\sqrt{N}\times\sqrt{N}}$

Very GPU-friendly: offline phase and online phase consist of matrix multiplications

On modern processors, can process query at *memory bandwidth*

2 operations per database entry

Query and response size:
$\sqrt{N}$ elements over $\mathbb{Z}_q$

PIR with *Silent* Preprocessing

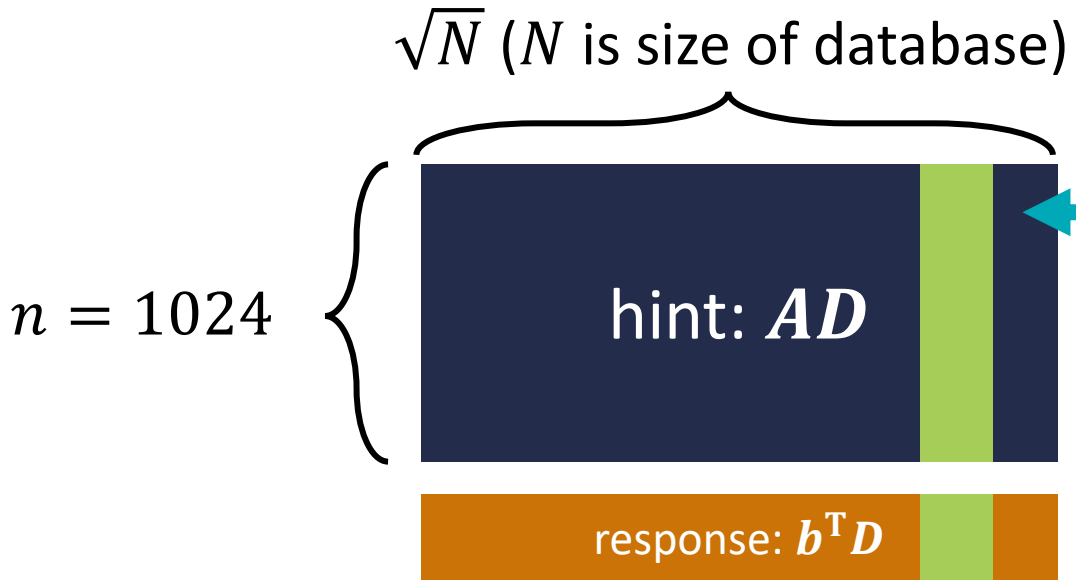
Silent preprocessing: allow **offline** preprocessing, but **no communication**

No need to manage hints; better suited for dynamic databases

Approach: *compress* the hint and include as part of the response

Why is the SimplePIR hint so large?

[Tiptoe; HDCZ23]
[HintlessPIR; LMRS24]
[YPIR; MW24]
[InsPIRe; MPSY26]



j^{th} column of hint is the "randomness" for the ciphertext that encrypts the j^{th} record

Hint is large because Regev encryption has poor rate: encrypting one scalar requires $(n+1)$ $\mathbb{Z}_q$ elements

Higher Rate using Polynomial Rings

Common technique to get better rate: use polynomial rings

Ciphertext space:

$$\mathbb{Z}_q^{n+1} \longrightarrow R_q^2 = \left(\mathbb{Z}_q[x]/(x^n+1)\right)^2$$

degree-n polynomials with coefficients in $\mathbb{Z}_q$

Each ciphertext now encrypts a degree-n polynomial (with coefficients in $\mathbb{Z}_p$)

$$\text{rate} = \frac{\text{size of plaintext}}{\text{size of ciphertext}} = \frac{\log p}{(n+1)\log q} \longrightarrow \frac{n \log p}{2n \log q} = \frac{\log p}{2 \log q}$$

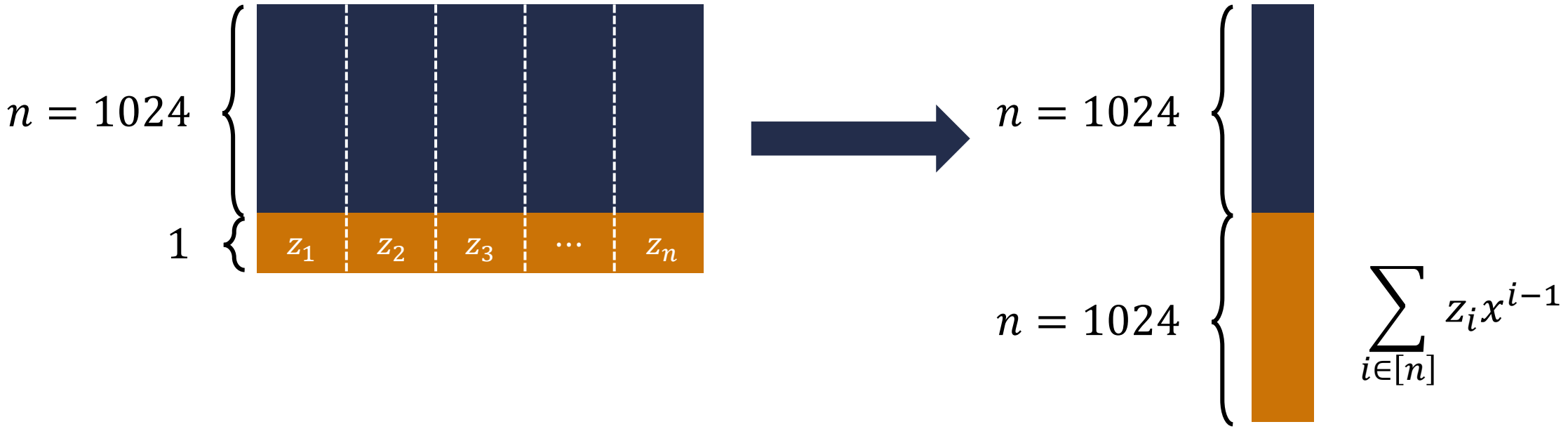
*rings allow **packing** more data into each ciphertext*

Recall $n = 2^{10} = 1024$, so using rings gives over 500× reduction in size

YPIR/InsPIRe: Packing LWE Ciphertexts

[MW24, MPSY26]

Approach: pack LWE ciphertexts into RLWE ciphertexts



$n(n+1)$ $\mathbb{Z}_q$ elements → $2n$ $\mathbb{Z}_q$ elements

over 500× reduction

Overall Protocol Flow

Compute:

Database $\boldsymbol{D} \in \mathbb{Z}_p^{\sqrt{N} \times \sqrt{N}}$

$\boldsymbol{AD}$

query packing components/matrices



No communication!

offline phase

online phase



$\boldsymbol{b}^{\mathrm{T}}, \widetilde{\boldsymbol{b}}_{\mathrm{ks},5}^{\mathrm{T}}, \widetilde{\boldsymbol{b}}_{\mathrm{ks},2n-1}^{\mathrm{T}}$



Limit: processing throughput is still constrained by memory bandwidth

Beyond Memory Bandwidth

[LG15, MW24, LHC25]

Can we go beyond memory bandwidth with a linear-scan PIR?

Yes, but in an ***amortized*** sense

Idea: if reading memory is the bottleneck, then schedule more CPU operations per byte of memory read

Cross-client batching: process a batch of k queries in parallel (from different, *non-coordinating* clients)

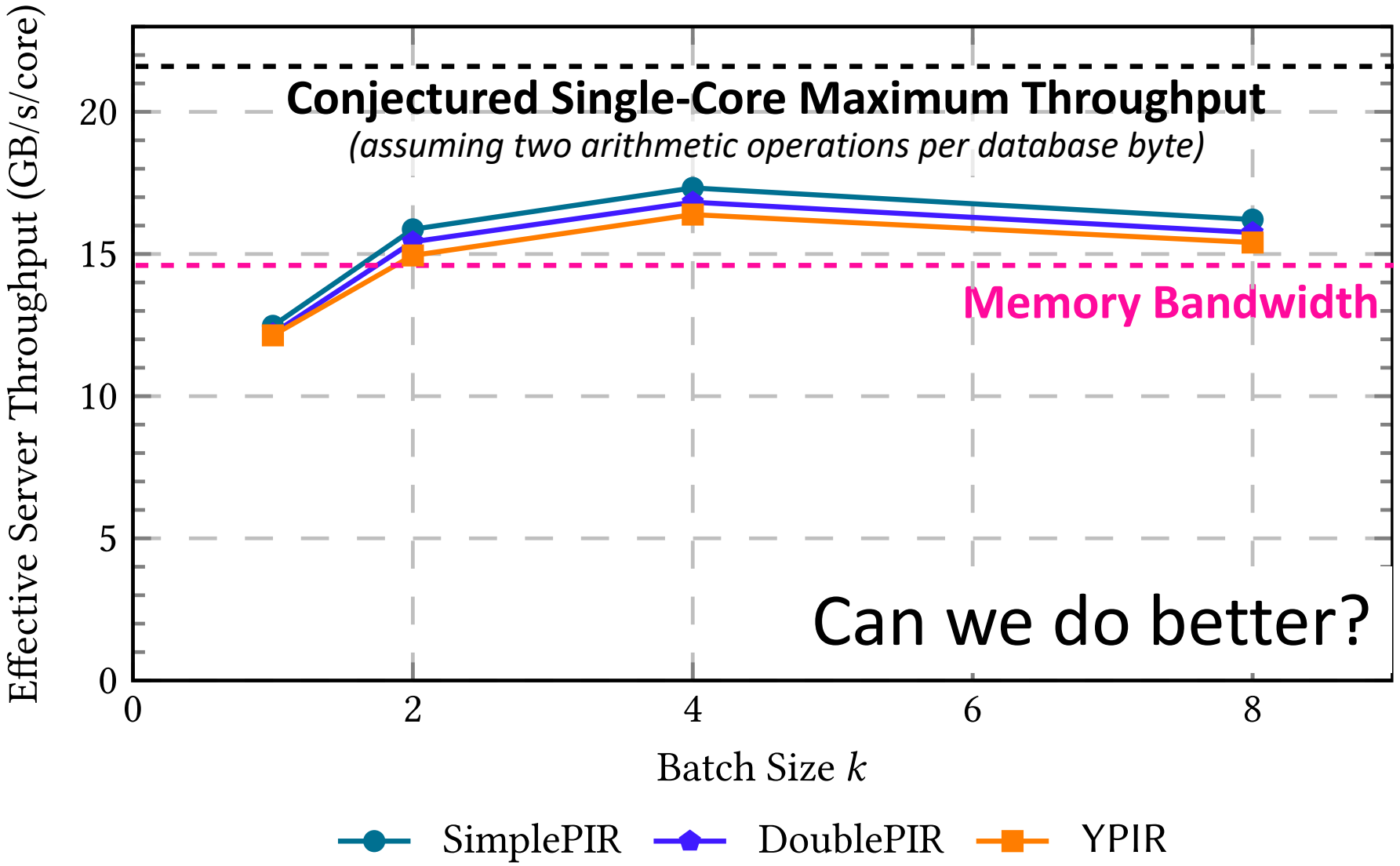
In many recent lattice-based PIR protocols, query processing is a product between a query vector and a database matrix

Cross-client batching: matrix-vector product ⇒ matrix-matrix product

Beyond Memory Bandwidth

[MW24]

$$\frac{k \cdot |\text{DB}|}{\text{Server time}}$$



Hardware Acceleration for PIR

Main computation in PIR: matrix multiplication



NVIDIA L40S

Memory bandwidth: 864 GB/s

64-core CPU: 1.4 trillion FLOPS

Computation: 91.6 trillion 32-bit (floating point) operations/s

Tensor cores: special architecture tailored for **low-precision** matrix multiplication operations for machine learning

INT8-INT32 tensor cores: takes two input matrices $\boldsymbol{A}, \boldsymbol{B}$ over $\mathbb{Z}_{2^8}$ and outputs product $\boldsymbol{AB}$ over $\mathbb{Z}_{2^{32}}$ $(\mathbb{Z}_{2^8} \times \mathbb{Z}_{2^8} \to \mathbb{Z}_{2^{32}})$

Throughput: 733 trillion INT-8 operations/s

"tensor core model of computation"

recent Intel CPUs also support it (Intel AMX)

Hardware Acceleration for PIR

Main computation in PIR: matrix multiplication



NVIDIA L40S

Memory bandwidth: 864 GB/s

64-core CPU: 1.4 trillion FLOPS

Computation: 91.6 trillion 32-bit (floating point) operations/s

Tensor cores: special architecture tailored for **low-precision** matrix multiplication operations for machine learning

INT8-INT32 tensor cores: takes two input matrices $\boldsymbol{A}, \boldsymbol{B}$ over $\mathbb{Z}_{2^8}$ and outputs product $\boldsymbol{AB}$ over $\mathbb{Z}_{2^{32}}$ $(\mathbb{Z}_{2^8} \times \mathbb{Z}_{2^8} \to \mathbb{Z}_{2^{32}})$

Throughput: 733 trillion INT-8 operations/s

Can we leverage tensor cores to accelerate PIR?

Performance Comparison

Setup: 4 GB database, 32 KB records

Implementation from [LHC25]

		CPU-based		GPU-based	
		HintlessPIR	**InsPIRe**	**SimplePIR**	**SandwichPIR**
Offline	Download	–	–	144 MB	–
Online	Upload	722 KB	1.04 MB	512 KB	576 KB
	Download	2.97 MB	96 KB	40 KB	112 KB
	Response Time	1.4 s	718 ms	33 ms	8.2 ms
	Throughput	2.8 GB/s	5.6 GB/s	123 GB/s	488 GB/s
	Cost (per million queries)	$489.89	$114.39	$20.49	$14.56

Excludes hint download costs

Performance Comparison

Setup: 4 GB database, 32 KB records

Implementation from [LHC25]

		CPU-based		GPU-based	
		HintlessPIR	**InsPIRe**	**SimplePIR**	
Offline	Download	–	–	144 MB	
Online	Upload			512 KB	
	Download			40 KB	112 KB
	Response Time			33 ms	8.2 ms
	Throughput	2.8 GB/s	5.6 GB/s	123 GB/s	488 GB/s
	Cost (per million queries)	$489.89	$114.39	$20.49	$14.56

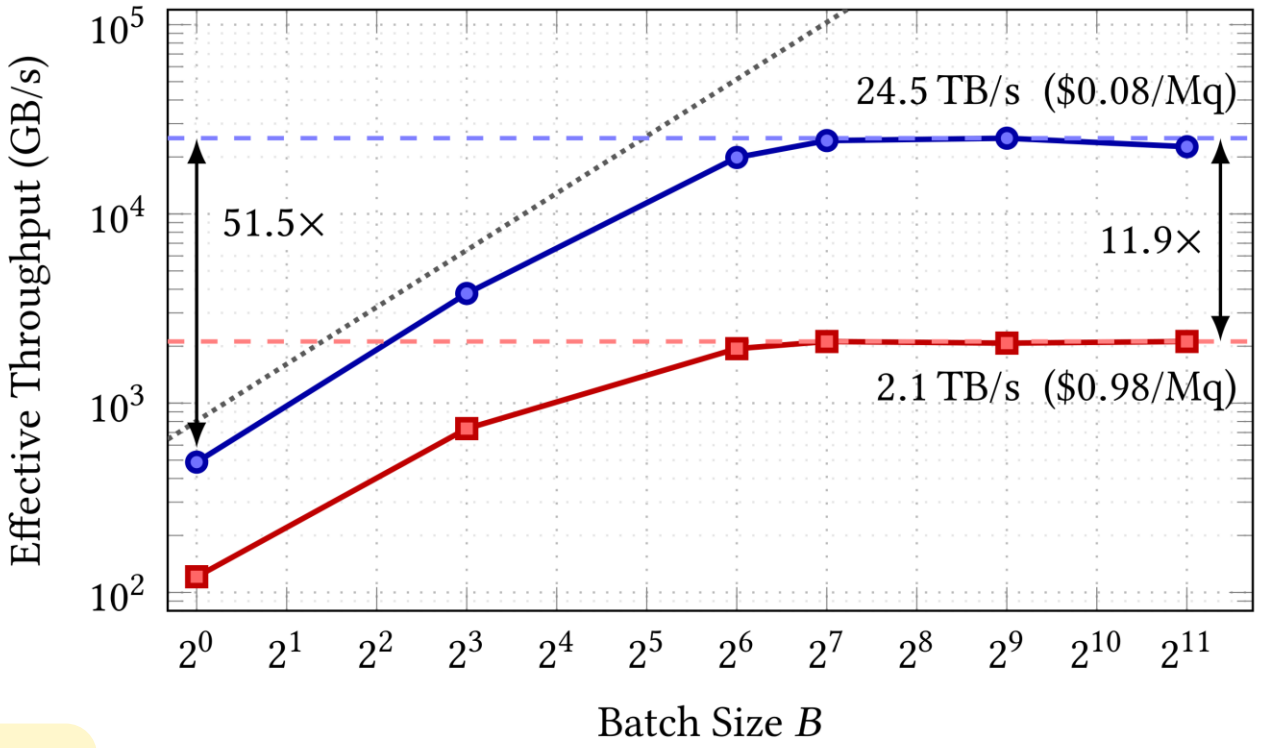
Over 88× higher throughput than CPU-based implementations for single-query setting (but we are not compute-bound yet)

72 KB is actually fixed (can be reused across queries); reduces per-query cost to $7.93

Excludes hint download costs

Cross-Client Batching

Cross-client batching: process multiple independent queries at once (more compute saturation)



$$\frac{k \cdot |\mathrm{DB}|}{\text{Server time}}$$

Effective throughput peaks around 25 TB/s

On typical CPUs, we expect ≈ 25 GB/s/core

(could be faster with special instruction sets)

Setup: 4 GB database, 32 KB records

Application to Private Wikipedia

Goal: read (text-only) Wikipedia over PIR

As of November 2023, English Wikipedia contains $6.4 \cdot 10^6$ articles
(8 GB after compression)

Record size is 128 KB (compressed articles are all under 128 KB)

Communication: 320 KB upload, 160 KB download
Server processing: $\approx$ 60 ms



https://www.cs.utexas.edu/~dwu4/pir-demo

Hardware Acceleration for Cryptography

Modern hardware is shaped by machine learning applications

For machine learning applications, elementary operation is matrix multiplication



In addition to Boolean circuits, arithmetic circuits, RAM programs, we should also think about "matrix multiplication programs"

Modern hardware designed for "matrix multiplication programs"

For PIR: using GPU tensor cores, we can now obtain (effective) throughputs of 25-275 TB/s (275 TB/s uses fleet of 8 GPUs over a 256 GB database)

Thank You!