

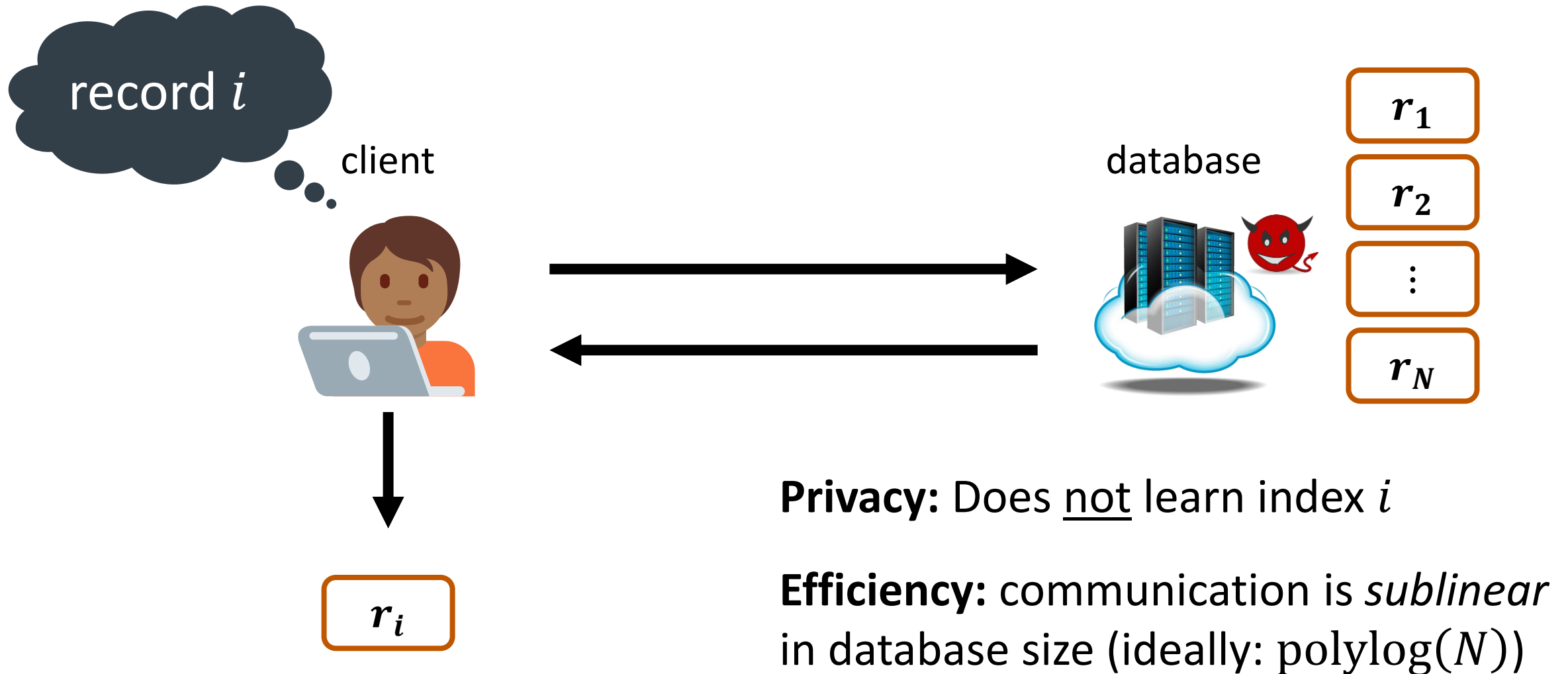
Recent Advancements in Private Information Retrieval

David Wu

based on joint works with Alexander Burton and Samir Menon

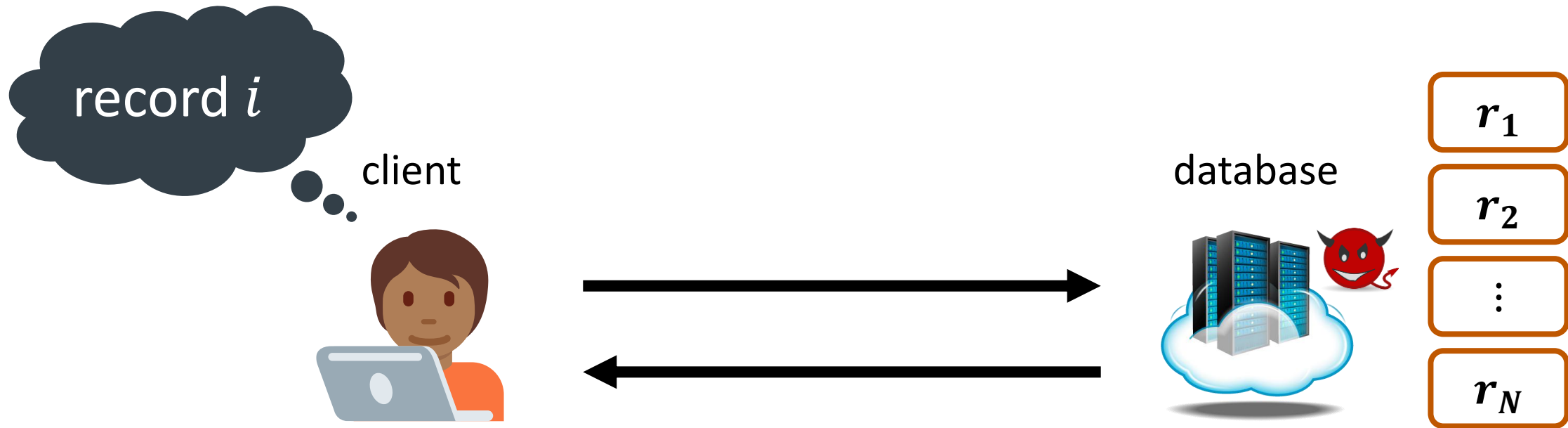
Private Information Retrieval (PIR)

[CGKS95]



Private Information Retrieval (PIR)

[CGKS95]



Basic building block in many privacy-preserving protocols

 Metadata-private messaging

 Contact discovery

 Private contact tracing

 Certificate transparency auditing

 Private web search

 Private DNS

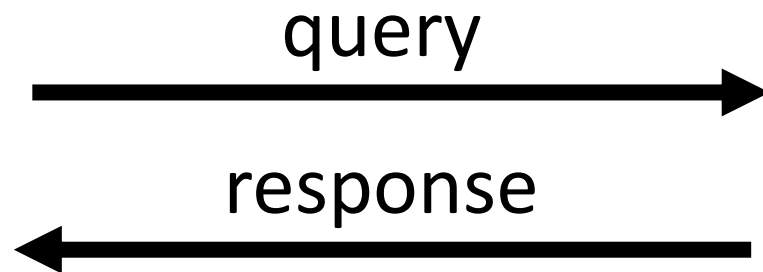
 Private content delivery

 Private navigation

 Password breach checking

Efficiency Metrics

1 Query size



2 Server Throughput

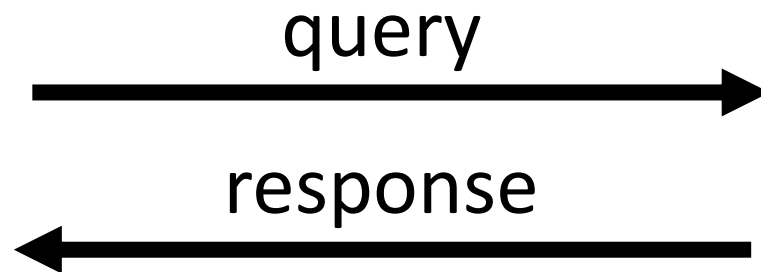
$$\frac{\text{database size}}{\text{server computation time}}$$

“measures how fast the server can respond as a function of database size”

Efficiency Metrics



1 Query size



Without preprocessing, server must perform a linear scan over the database

2 Server Throughput

$$\frac{\text{database size}}{\text{server computation time}}$$

“measures how fast the server can respond as a function of database size”

Efficiency Metrics

Client generates a *reusable* set of public parameters



public parameters
→

1 Query size

query
→

←
response

3 Rate

$$\frac{\text{record size}}{\text{response size}}$$

“measures communication overhead in responses”

4 Public parameter size

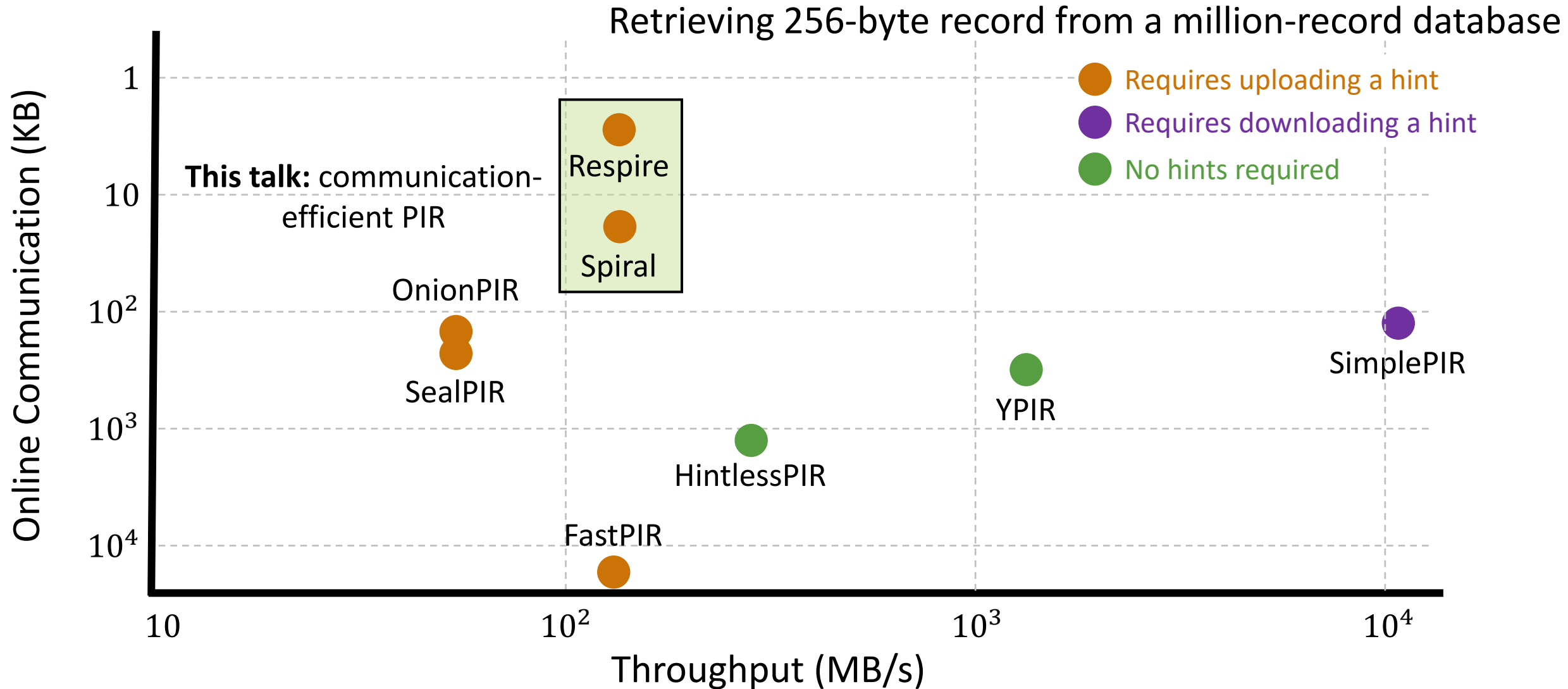


2 Server Throughput

$$\frac{\text{database size}}{\text{server computation time}}$$

“measures how fast the server can respond as a function of database size”

Communication/Computation Trade-offs in PIR



PIR from Homomorphic Encryption

[K097]

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

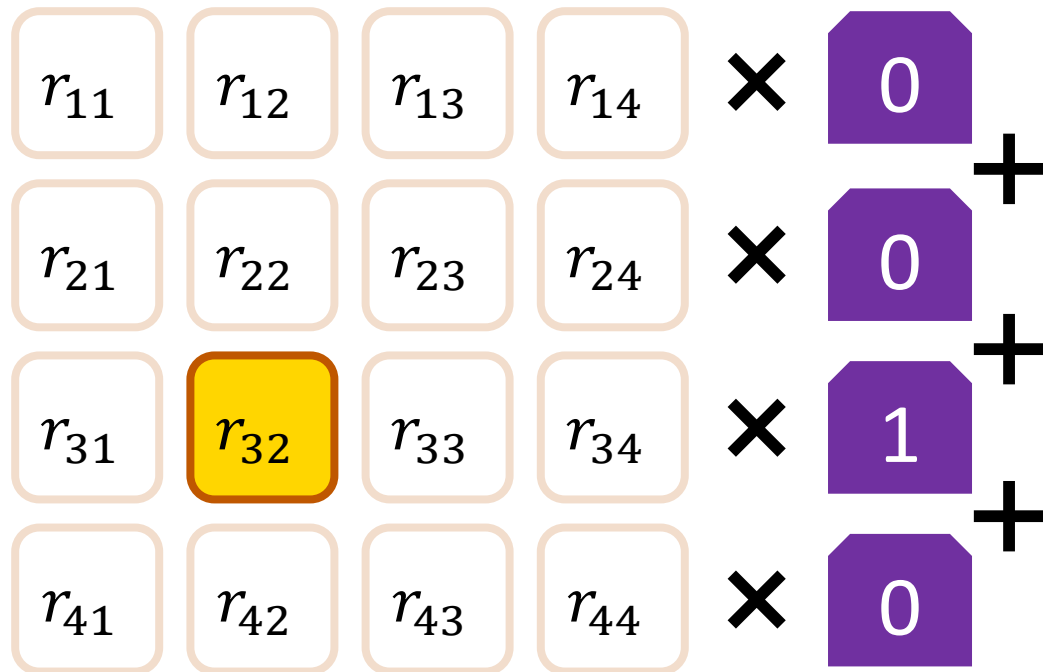
r_{11}	r_{12}	r_{13}	r_{14}
r_{21}	r_{22}	r_{23}	r_{24}
r_{31}	r_{32}	r_{33}	r_{34}
r_{41}	r_{42}	r_{43}	r_{44}

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

PIR from Homomorphic Encryption

[K097]

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



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

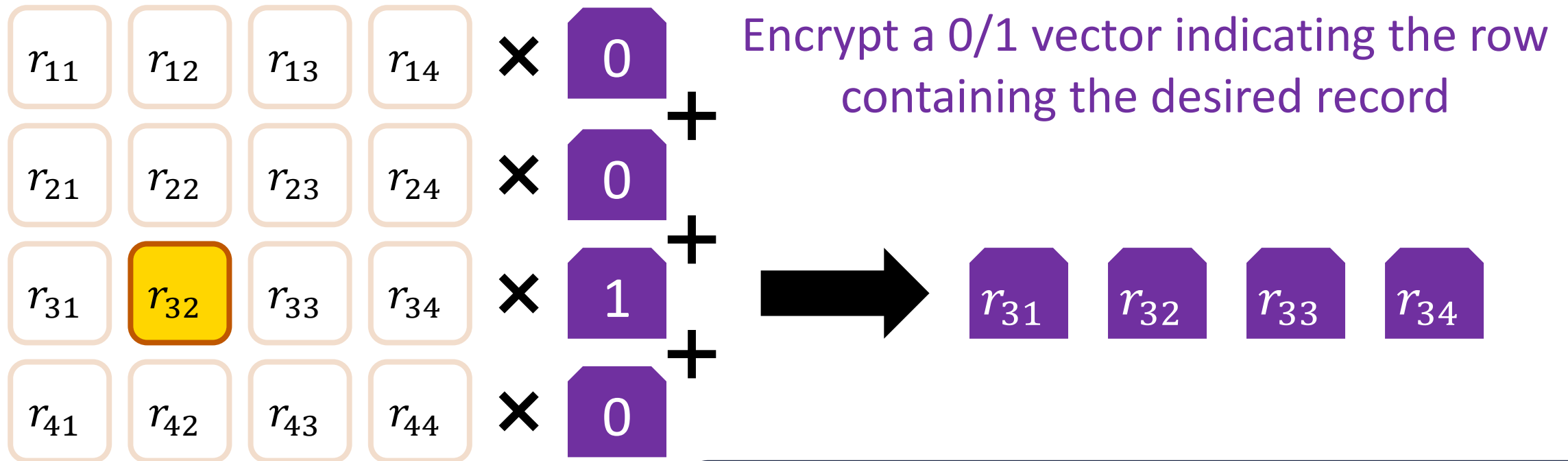
Arrange the database as a \sqrt{N} -by- \sqrt{N} matrix

Homomorphically compute product between query vector and database matrix

PIR from Homomorphic Encryption

[K097]

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



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

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

PIR from Homomorphic Encryption

[K097]

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



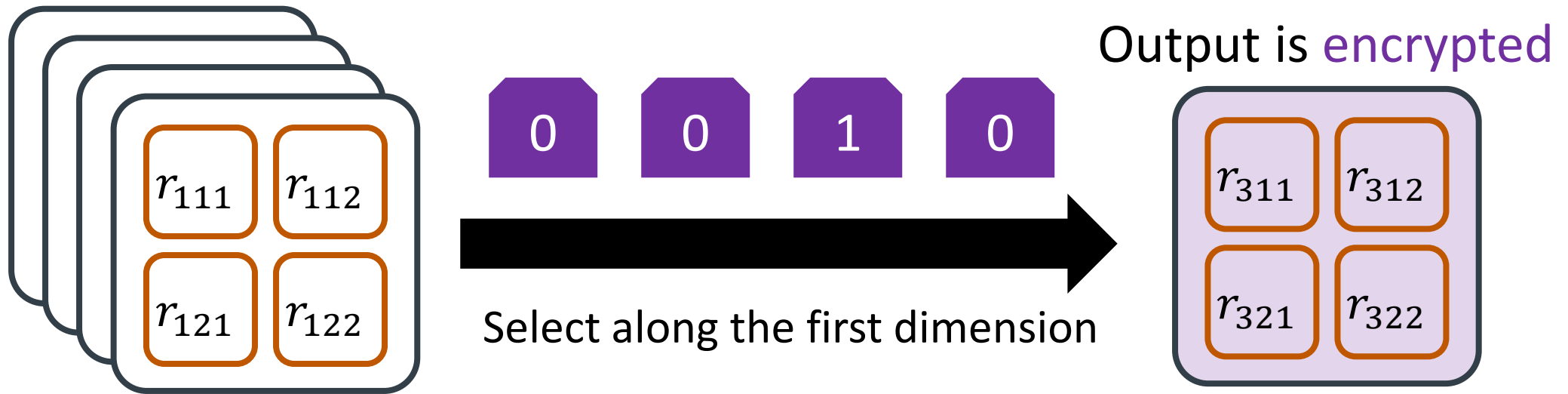
Response size: $O_\lambda(\sqrt{N})$

Homomorphically compute product
between query vector and database matrix

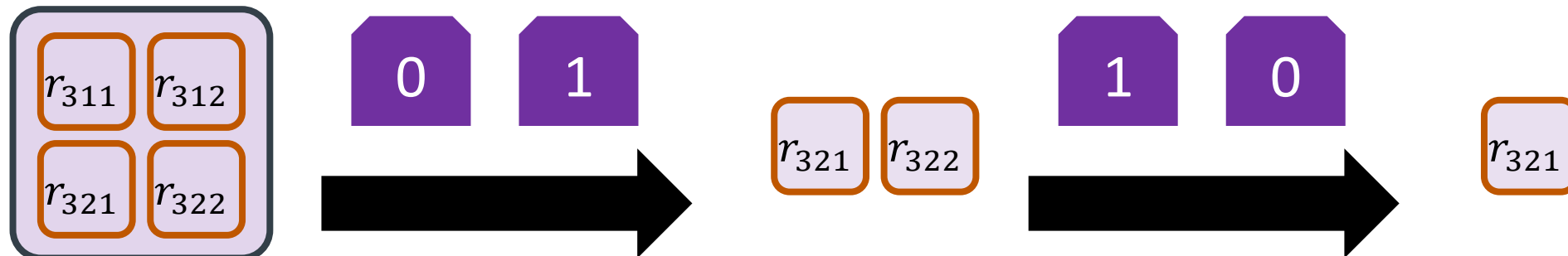
PIR from Homomorphic Encryption

[K097]

Sub- \sqrt{N} communication: view the database as **hypercube**



Approach: Use **homomorphic multiplication** [GH19, PT20, ALPRSSY21, MCR21]



SPIRAL: Composing FHE Schemes

Follows Gentry-Halevi blueprint of composing **two** lattice-based encryption schemes:

Ciphertexts in lattice-based schemes are noisy encodings

Homomorphic operations increase noise; more noise = larger parameters = less efficiency

Scheme 1: Regev's encryption scheme [Reg04]

Small ciphertexts (amortized); only supports additive homomorphism

18 KB plaintext \Rightarrow 43 KB ciphertext (2.4 \times expansion)

1 MB plaintext \Rightarrow 1.3 MB ciphertext (1.3 \times expansion)

allows the use of
smaller lattice
dimension and modulus

Scheme 2: Gentry-Sahai-Waters encryption scheme [GSW13]

Large ciphertexts; supports homomorphic multiplication (with additive noise growth)

1 bit plaintext \Rightarrow 2.5 **MB** ciphertext

Can we get the best of both worlds?

SPIRAL: Composing FHE Schemes

Follows Gentry-Halevi blueprint of composing **two** lattice-based encryption schemes:

Ciphertexts in lattice-based schemes are noisy encodings

Homomorphic operations increase noise; more noise = larger parameters = less efficiency

Scheme 1: Regev's encryption scheme [Reg04]

Small ciphertexts (amortized); only supports additive homomorphism

18 KB plaintext \Rightarrow 43 KB ciphertext (2.4 \times expansion)

1 MB plaintext \Rightarrow 1.3 MB ciphertext (1.3 \times expansion)

allows the use of
smaller lattice
dimension and modulus

Scheme 2: Gentry-Sahai-Waters encryption scheme [GSW13]

Large ciphertexts; supports homomorphic multiplication (with additive noise growth)

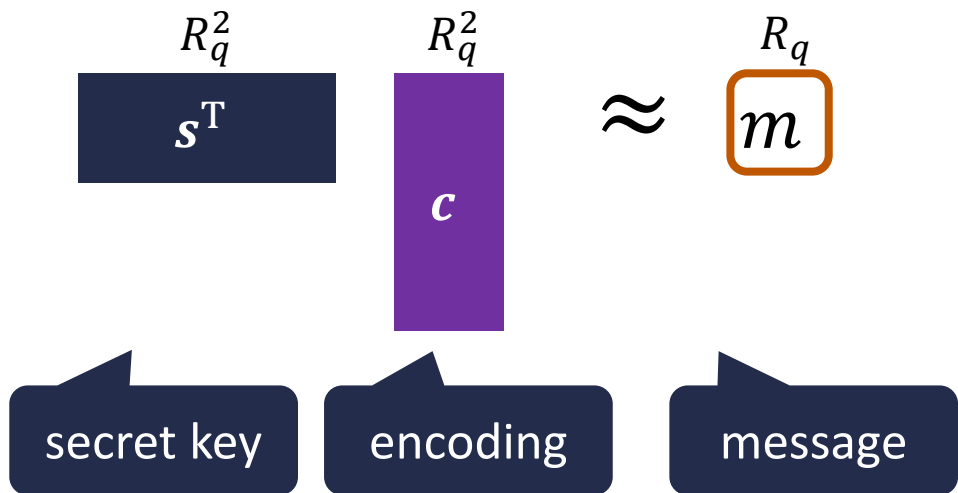
1 bit plaintext \Rightarrow 2.5 **MB** ciphertext

SPIRAL: Use GSW for homomorphic multiplication, Regev for communication

Regev Encodings (over Rings)

[Reg04, LPR10]

Regev encoding of a scalar $m \in R$:



- Secret key allows recovery of **noisy** version of original message
- To support decryption of “small” values $t \in R_p$, we encode t as $(q/p)t$
- Decryption recovers noisy version of $(q/p)t$ and rounding yields t

$$\text{rate} = \frac{\log p}{2 \log q} < \frac{1}{2}$$

OnionPIR: rate = 0.24

All elements are polynomials in the ring $R = \mathbb{Z}[x]/(x^d + 1)$ where $d = 2^k$

Matrix Regev Encodings (over Rings)

[PVW08, LPR10]

Regev encoding of a matrix $M \in R_q^{n \times n}$:

Idea: “Reuse” encryption randomness

$$\begin{array}{ccc} R_q^{n \times (n+1)} & R_q^{(n+1) \times n} & R_q^{n \times n} \\ \boxed{S^T} & \boxed{C} & \approx \boxed{M} \end{array}$$

$$\text{rate} = \frac{n^2 \log p}{n(n+1) \log q} = \frac{n^2 \log p}{n^2 + n \log q}$$

Additively homomorphic:

$$S^T C_1 \approx M_1$$

$$S^T C_2 \approx M_2$$

$$S^T (C_1 + C_2) \approx M_1 + M_2$$

All elements are polynomials in the ring $R = \mathbb{Z}[x]/(x^d + 1)$ where $d = 2^k$

Gentry-Sahai-Waters Encodings

[GSW13]

GSW encoding of a bit $\mu \in \{0,1\}$:

$$\begin{array}{c}
 R_q^{n \times (n+1)} \quad R_q^{(n+1) \times n} \\
 \mathcal{S}^T \quad \mathcal{C} \approx \boxed{\mu} \quad \mathcal{S}^T \quad G \\
 R_q^{n \times (n+1)} \quad R_q^{(n+1) \times m}
 \end{array}$$

$m = (n + 1) \log q$

Gadget matrix [MP12]:

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}^T & & & & \\ & \ddots & & & \\ & & & & \mathbf{g}^T \end{bmatrix}$$

$$\mathbf{g}^T = [1 \quad 2 \quad 2^2 \quad \dots \quad 2^{\lfloor \log_2 q \rfloor}]$$

“Powers-of-2” matrix

Main property: for every vector $\mathbf{v} \in \mathbb{Z}_q^{n+1}$, can define $\mathbf{G}^{-1}(\mathbf{v}) \in \{0,1\}^m$ where $\mathbf{G}\mathbf{G}^{-1}(\mathbf{v}) = \mathbf{v}$
“binary decomposition”

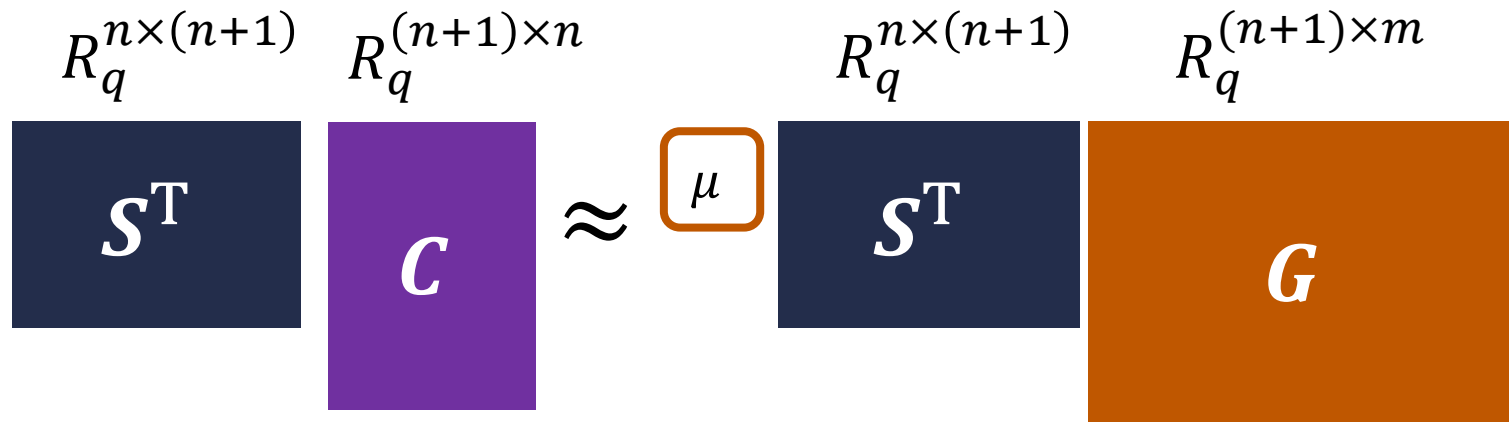
Construction will use other decomposition bases

All elements are polynomials in the ring $R = \mathbb{Z}[x]/(x^d + 1)$ where $d = 2^k$

Gentry-Sahai-Waters Encodings

[GSW13]

GSW encoding of a bit $\mu \in \{0,1\}$:



Gadget matrix [MP12]:

$$G = \begin{bmatrix} \mathbf{g}^T & & & & \\ & \ddots & & & \\ & & & & \\ & & & & \mathbf{g}^T \end{bmatrix}$$

$$\mathbf{g}^T = [1 \quad 2 \quad 2^2 \quad \dots \quad 2^{\lfloor \log_z q \rfloor}]$$

$$m = (n + 1) \log q$$

“Powers-of-2” matrix

$$\text{rate} = \frac{1}{d(n+1)^2 \log q}$$

Concretely: $d = 2048, n \geq 1, q = 2^{56}$

Construction will use other decomposition bases

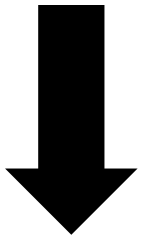
All elements are polynomials in the ring $R = \mathbb{Z}[x]/(x^d + 1)$ where $d = 2^k$

Regev-GSW Homomorphism

[CGGI18]

$$\mathbf{S}^T \mathbf{C}_{\text{Reg}} \approx \mathbf{M}$$

$$\mathbf{S}^T \mathbf{C}_{\text{GSW}} \approx \mu \mathbf{S}^T \mathbf{G}$$



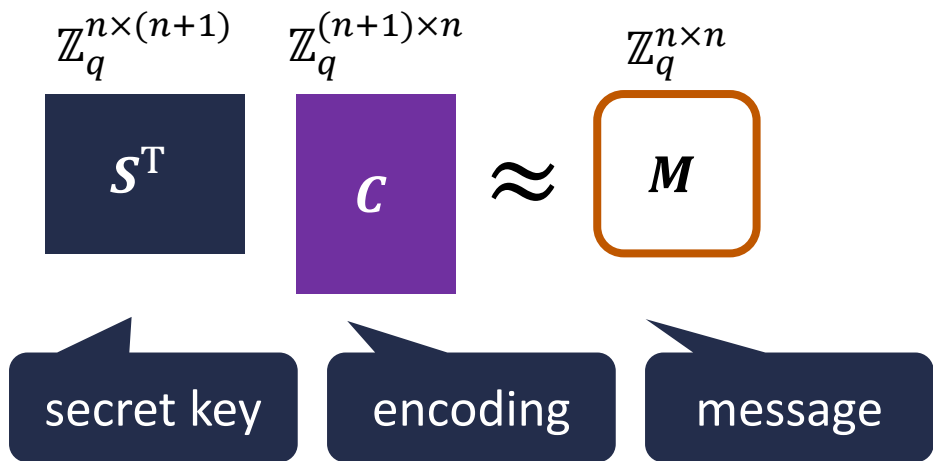
$$\mathbf{S}^T \mathbf{C}_{\text{GSW}} \mathbf{G}^{-1}(\mathbf{C}_{\text{Reg}}) \approx \mu \mathbf{S}^T \mathbf{C}_{\text{Reg}} \approx \mu \mathbf{M}$$

$\mathbf{C}_{\text{GSW}} \mathbf{G}^{-1}(\mathbf{C}_{\text{Reg}})$ is a Regev encoding of $\mu \mathbf{M}$

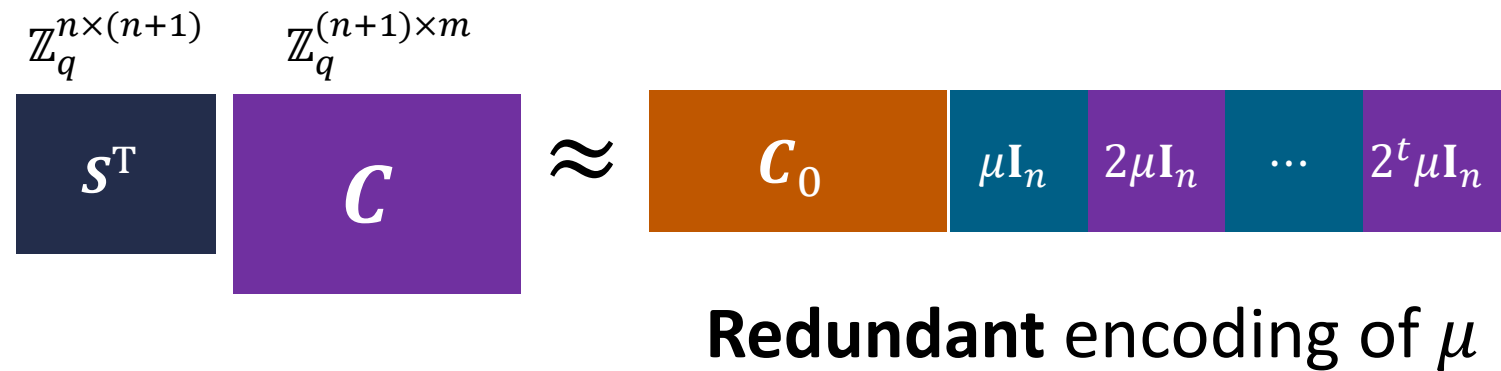
Regev-GSW Homomorphism

[CGGI18]

Regev encoding of $M \in \mathbb{Z}_q^{n \times n}$:



GSW encoding of $\mu \in \mathbb{Z}_q$:

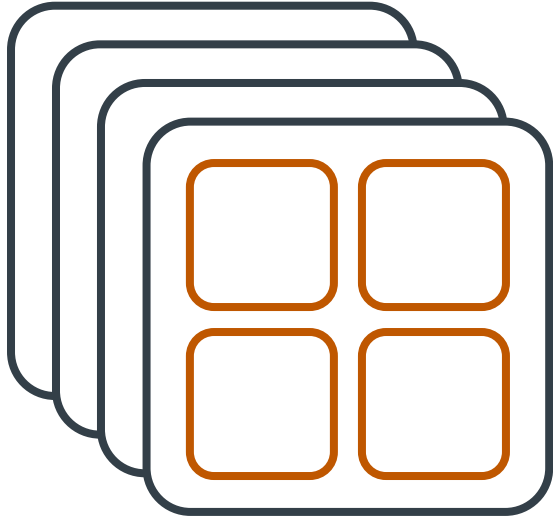


$$S^T = [-s \mid I_n] \in R_q^{n \times (n+1)}$$

Key property: given Regev encoding of message M and GSW encoding of scalar μ , can efficiently derive a Regev encoding of $\mu \cdot M$

The Gentry-Halevi Blueprint

[GH19]



Database is represented as $2^{\nu_1} \times \underbrace{2 \times 2 \times \dots \times 2}_{2^{\nu_2}}$ hypercube

Query contains 2^{ν_1} matrix Regev ciphertexts



Indicator for index along first dimension

Query contains ν_2 GSW ciphertexts



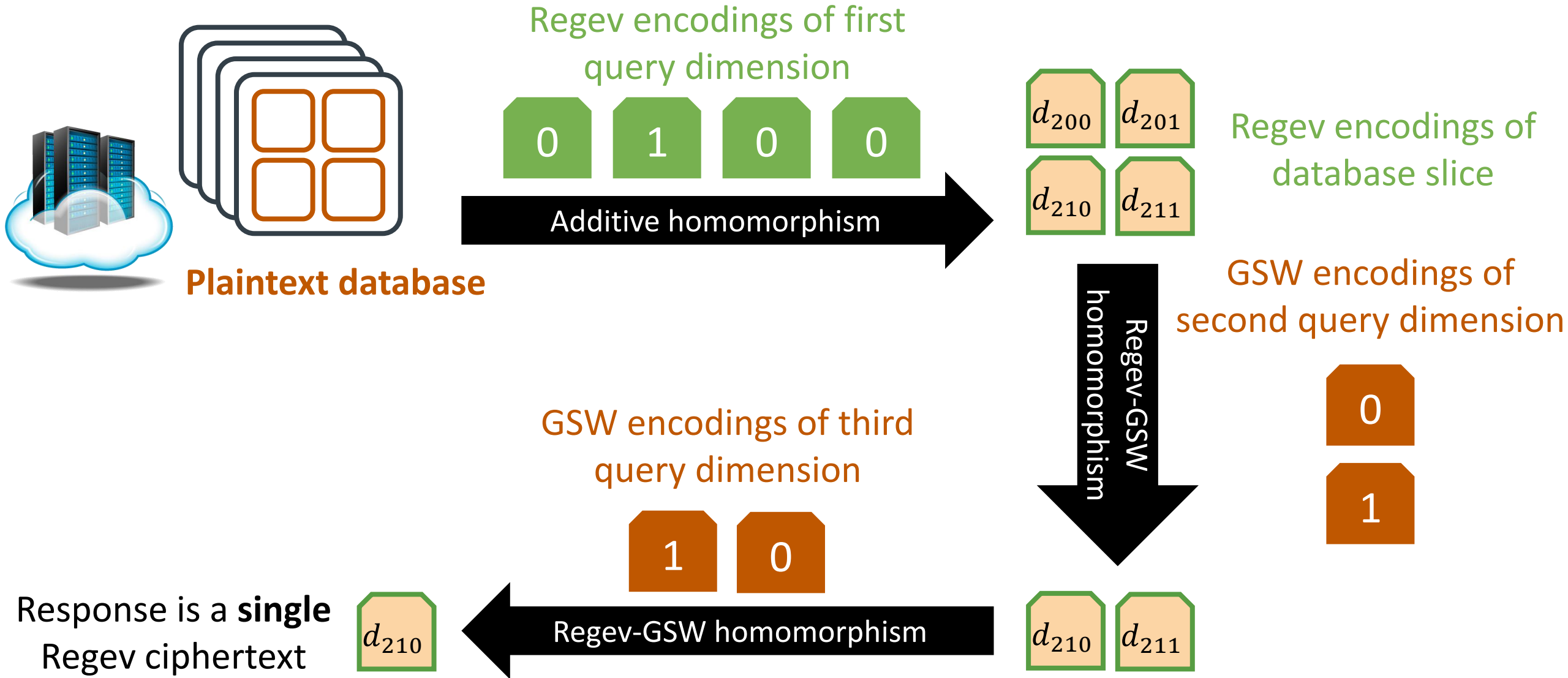
Indicator for index along subsequent dimensions

Response is a single matrix Regev ciphertext

Each GSW ciphertext participates in only one multiplication with a Regev ciphertext!

The Gentry-Halevi Blueprint

[GH19]



The Gentry-Halevi Blueprint

[GH19]

Database is represented as $2^{\nu_1} \times \underbrace{2 \times 2 \times \dots \times 2}_{2^{\nu_2}}$ hypercube

Drawback: large queries

Can compress using polynomial encoding method of Angel et al. [ACLS18]

Query contains 2^{ν_1} matrix Regev ciphertexts



Indicator for index along first dimension

Query contains ν_2 GSW ciphertexts



Indicator for index along subsequent dimensions

Estimated size:
4 MB/ciphertext

Estimated query size:
30 MB

The Gentry-Halevi Blueprint

[GH19]

Database is represented as $2^{\nu_1} \times \underbrace{2 \times 2 \times \dots \times 2}_{2^{\nu_2}}$ hypercube

Drawback: large queries

Can compress using polynomial encoding method of Angel et al. [ACLS18]

Query contains 2^{ν_1} matrix Regev ciphertexts



Indicator for index along first dimension

Query contains ν_2 GSW ciphertexts



Indicator for index along subsequent dimensions

SealPIR query size:
66 KB

Estimated query size:
30 MB

The SPIRAL Protocol

Key idea: Expand Regev encodings into GSW encodings

OnionPIR [MCR21]: use Regev-GSW homomorphism for the **scalar** case

SPIRAL: General toolkit to translate between Regev and GSW

Transformations useful for **query compression** and **response packing**

Assembling GSW Encodings

Goal: use **Regev** encodings to construct \mathbf{C} such that $\mathbf{S}^T \mathbf{C} \approx \mu \mathbf{S}^T \mathbf{G}$

$$\mu \mathbf{S}^T \mathbf{G} = \begin{array}{|c|c|c|c|c|c|} \hline \mathbf{C}_0 & \mu \mathbf{I}_n & 2\mu \mathbf{I}_n & 2^2 \mu \mathbf{I}_n & \dots & 2^t \mu \mathbf{I}_n \\ \hline \end{array}$$

$$\mathbf{C} = \begin{array}{|c|c|c|c|c|c|} \hline \mathbf{A} & \mathbf{B}_0 & \mathbf{B}_1 & \mathbf{B}_2 & \dots & \mathbf{B}_t \\ \hline \end{array}$$

Break \mathbf{C} into *blocks*

Query Compression in SPIRAL

Database is represented as $2^{\nu_1} \times \underbrace{2 \times 2 \times \dots \times 2}_{2^{\nu_2}}$ hypercube

Query contains 2^{ν_1} matrix Regev encodings



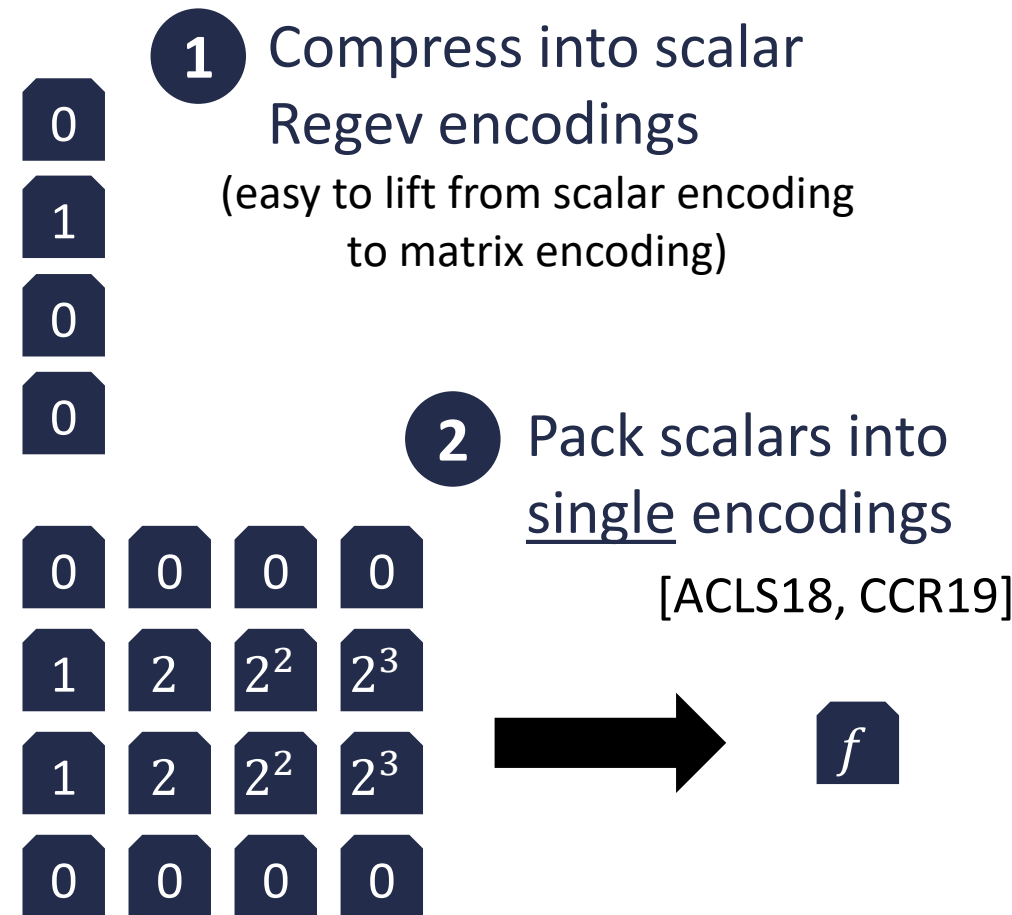
Indicator for index along first dimension

Query contains ν_2 GSW encodings



Indicator for index along subsequent dimensions

Similar techniques possible for *response* compression [see paper]



The SPIRAL Protocol

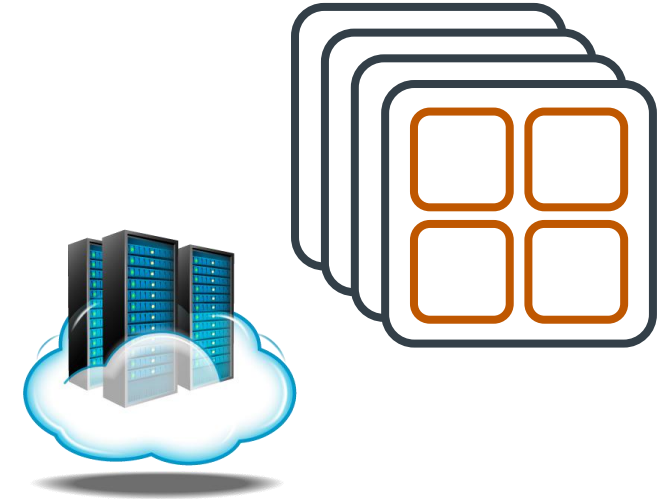
record i



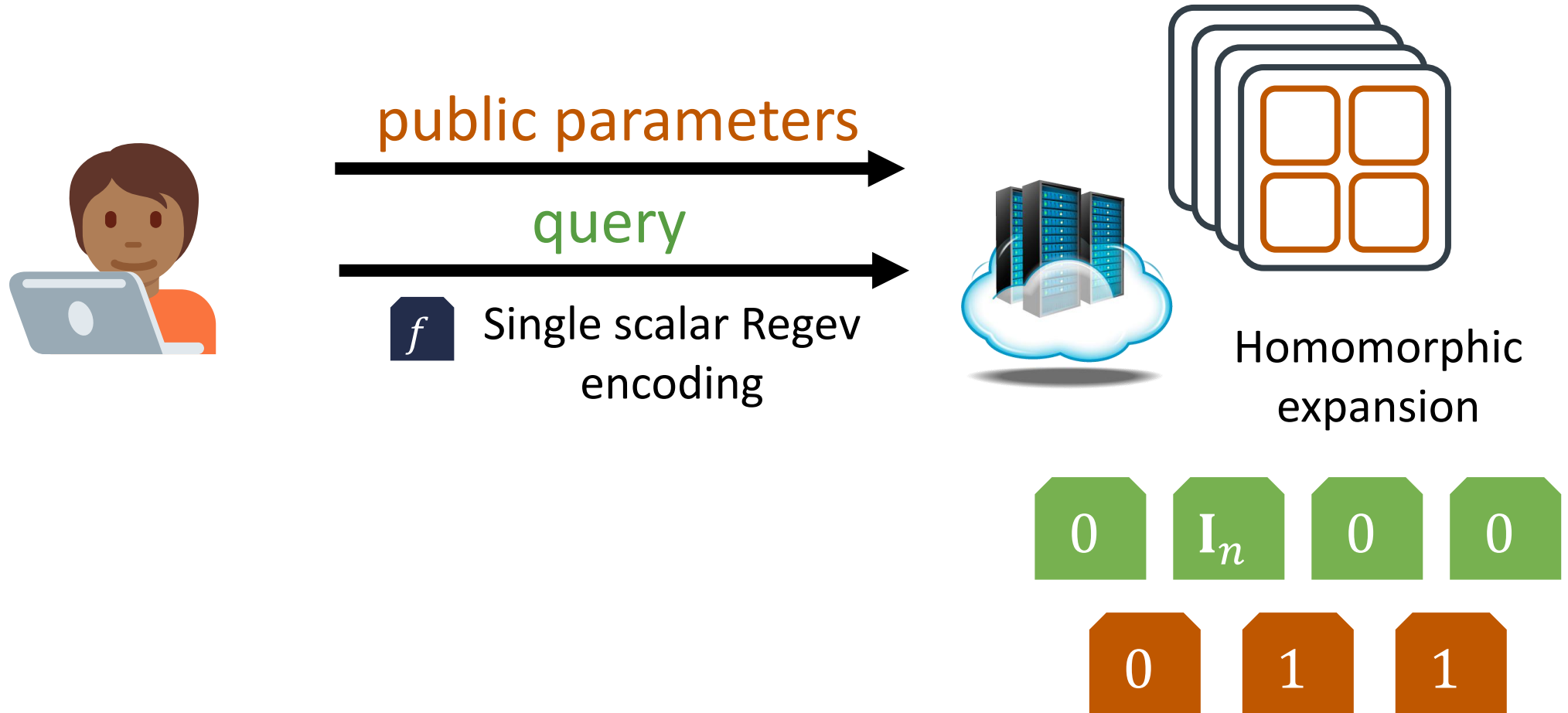
public parameters



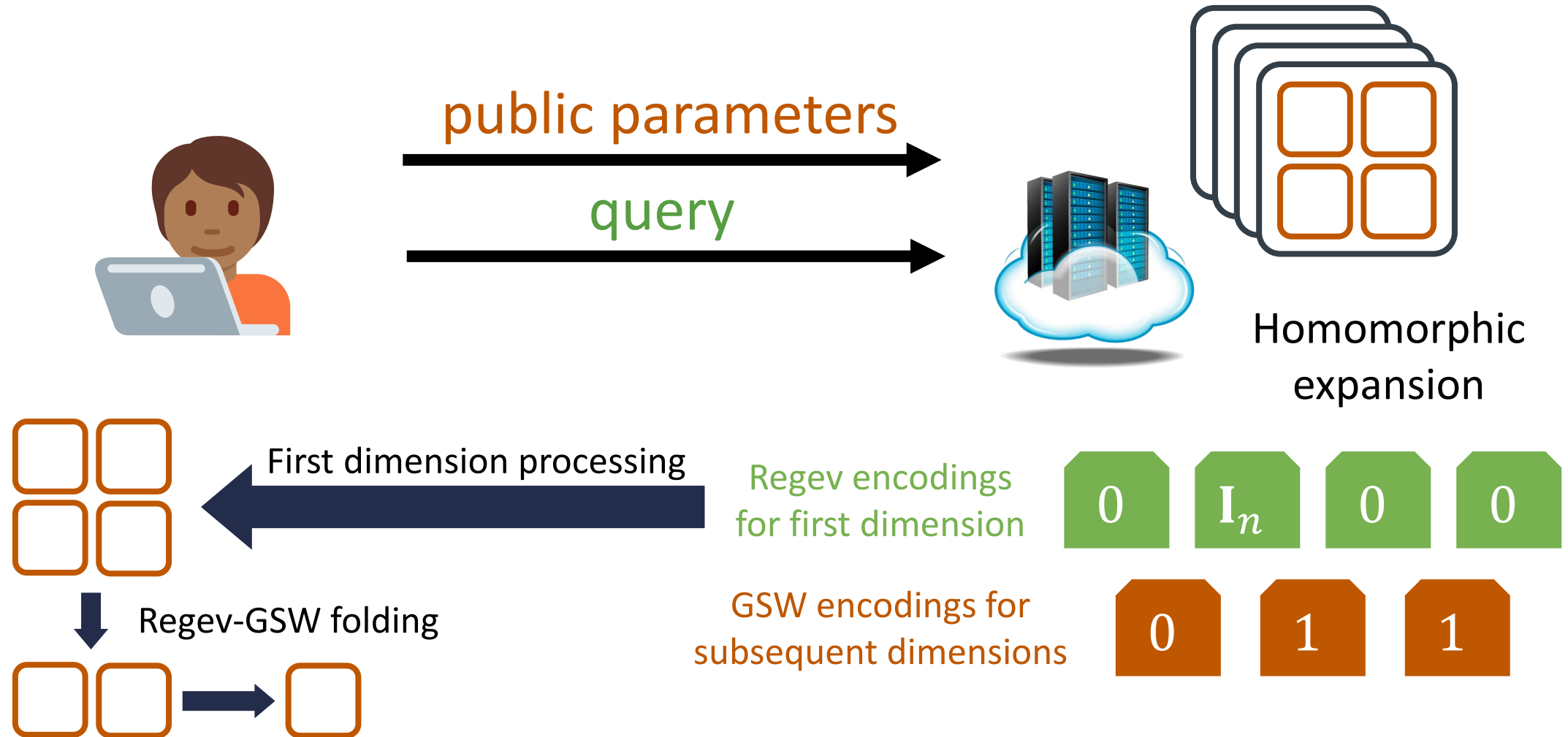
Key-switching matrices for
ciphertext expansion and
translation



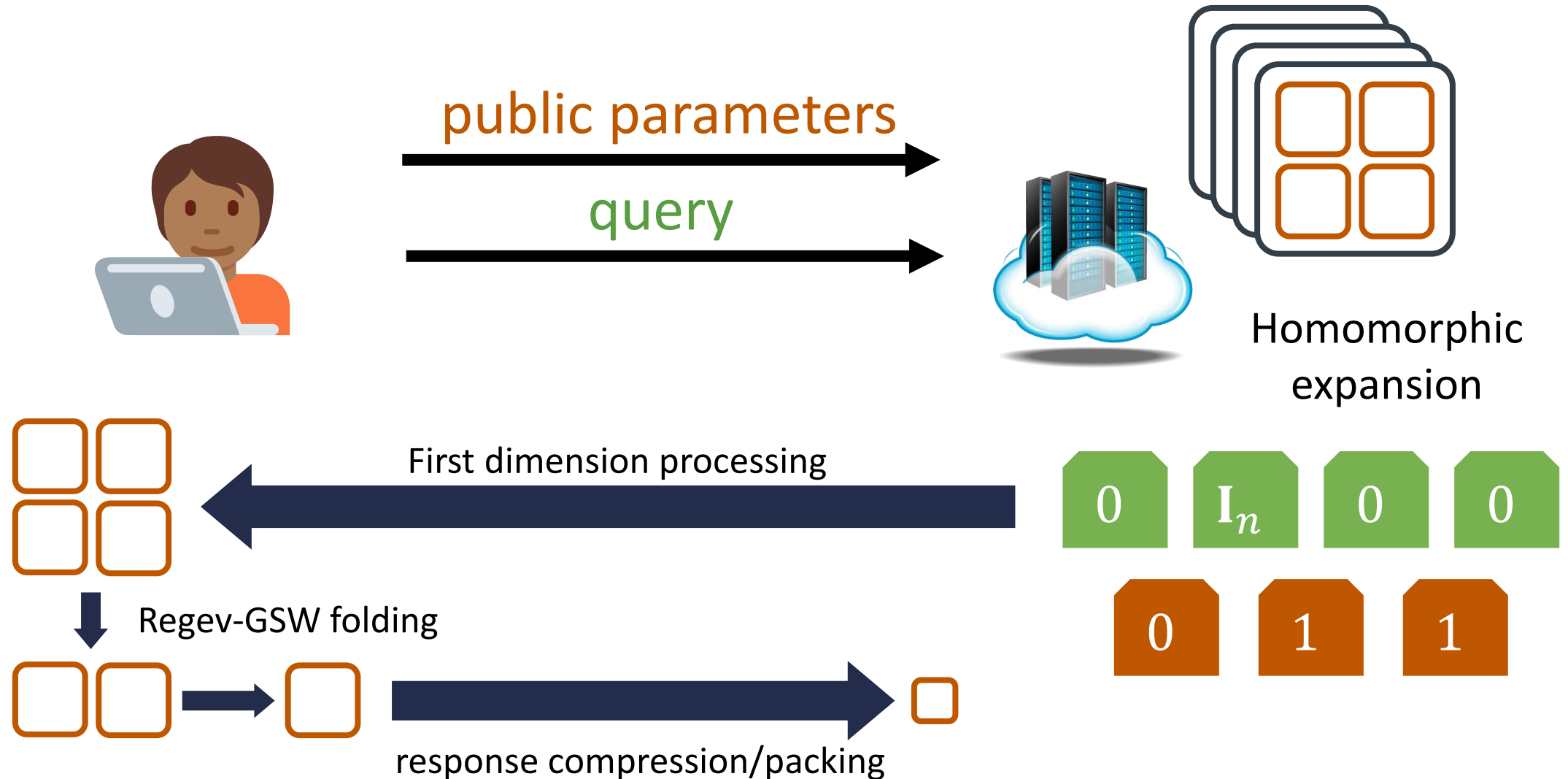
The SPIRAL Protocol



The SPIRAL Protocol



The SPIRAL Protocol



Basic Comparisons

Database	Metric	SealPIR	FastPIR	OnionPIR	SPIRAL
2¹⁸ records 30 KB records (7.9 GB database)	Public Param. Size	3 MB	1 MB	5 MB	18 MB
	Query Size	66 KB	8 MB	63 KB	14 KB
	Response Size	3 MB	262 KB	127 KB	84 KB
	Server Compute	74.91 s	50.5 s	52.7 s	24.5 s
			Rate:	0.24	0.36
			Throughput:	149 MB/s	322 MB/s

Database configuration preferred by OnionPIR

Compared to OnionPIR:

reduce query size by 4.5×

increase public parameter size by 3.6×

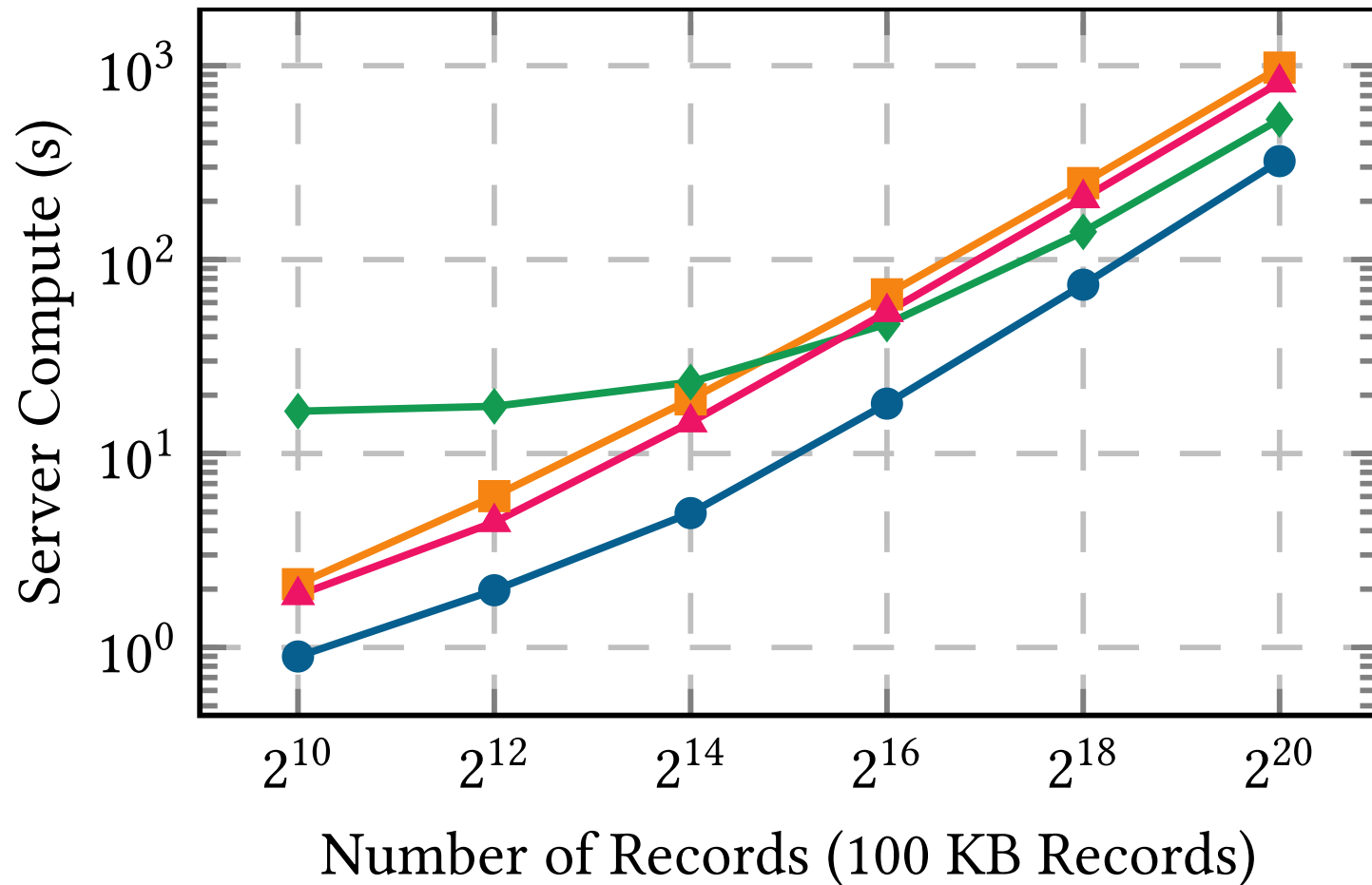
reduce response size by 2×

reduce compute time by 2×

Comparisons against other communication-efficient schemes (i.e., ones that do not have server hints)

In particular, these exclude subsequent schemes such as FrodoPIR, SimplePIR, and Piano

Basic Comparisons (with Large Records)



Throughput for 100 GB database (2^{20} records):

- SPIRAL: 310 MB/s (322 s)
- SealPIR: 102 MB/s (977 s)
- FastPIR: 189 MB/s (528 s)
- OnionPIR: 122 MB/s (817 s)

SPIRAL also has smaller query size and response size, but larger public parameters

All measurements based on single-thread/single-core processing

—●— SPIRAL —■— SealPIR —◆— FastPIR —▲— OnionPIR

The Streaming Setting

Streaming setting: same query reused over multiple databases

Private video stream (database D_i contains i^{th} block of media) [GCMSAW16]

Private voice calls (repeated polling of the same “mailbox”) [AS16, AYAAG21]

Goal: minimize online costs (i.e., server compute, response size)

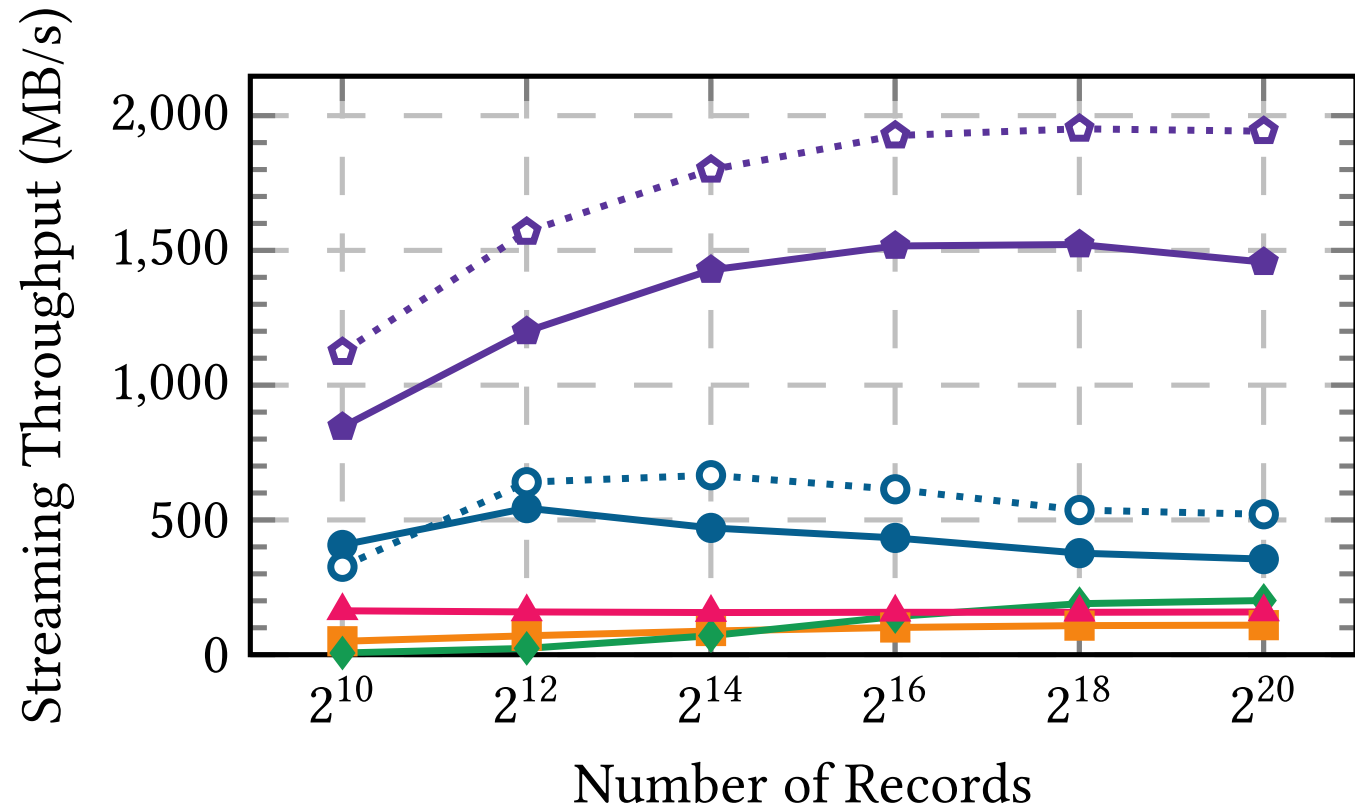
Consider larger public parameters or query size (amortized over lifetime of stream)

Approach: send all of the Regev encodings (and only use Regev-GSW translation)

The Streaming Setting

Streaming throughput: ignoring query expansion costs, assuming optimal record size for each system

Packing outperforms non-packed protocol for streaming settings

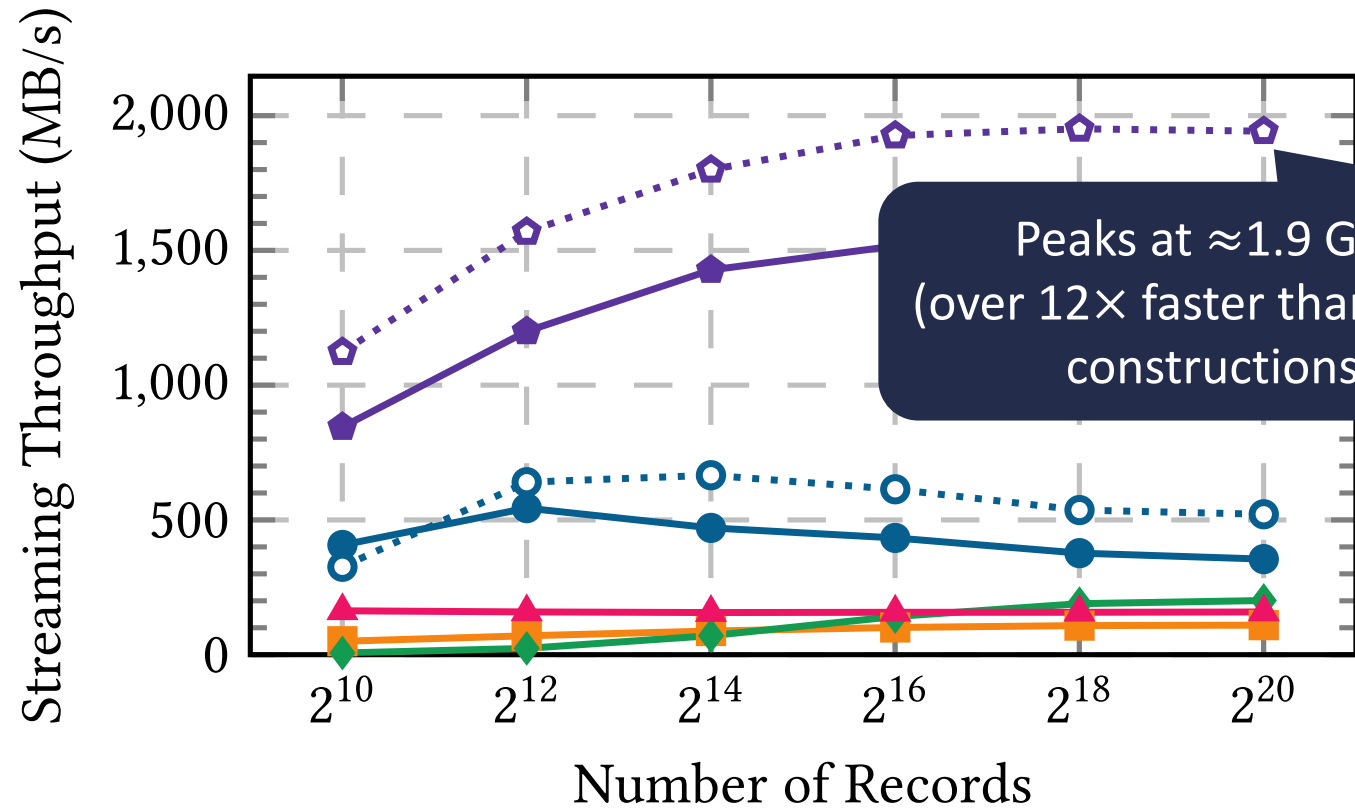


Packed versions rely on response compression (larger public parameters, higher throughput)



The Streaming Setting

Streaming throughput: ignoring query expansion costs, assuming optimal record size for each system



Packing outperforms non-packed protocol for streaming settings

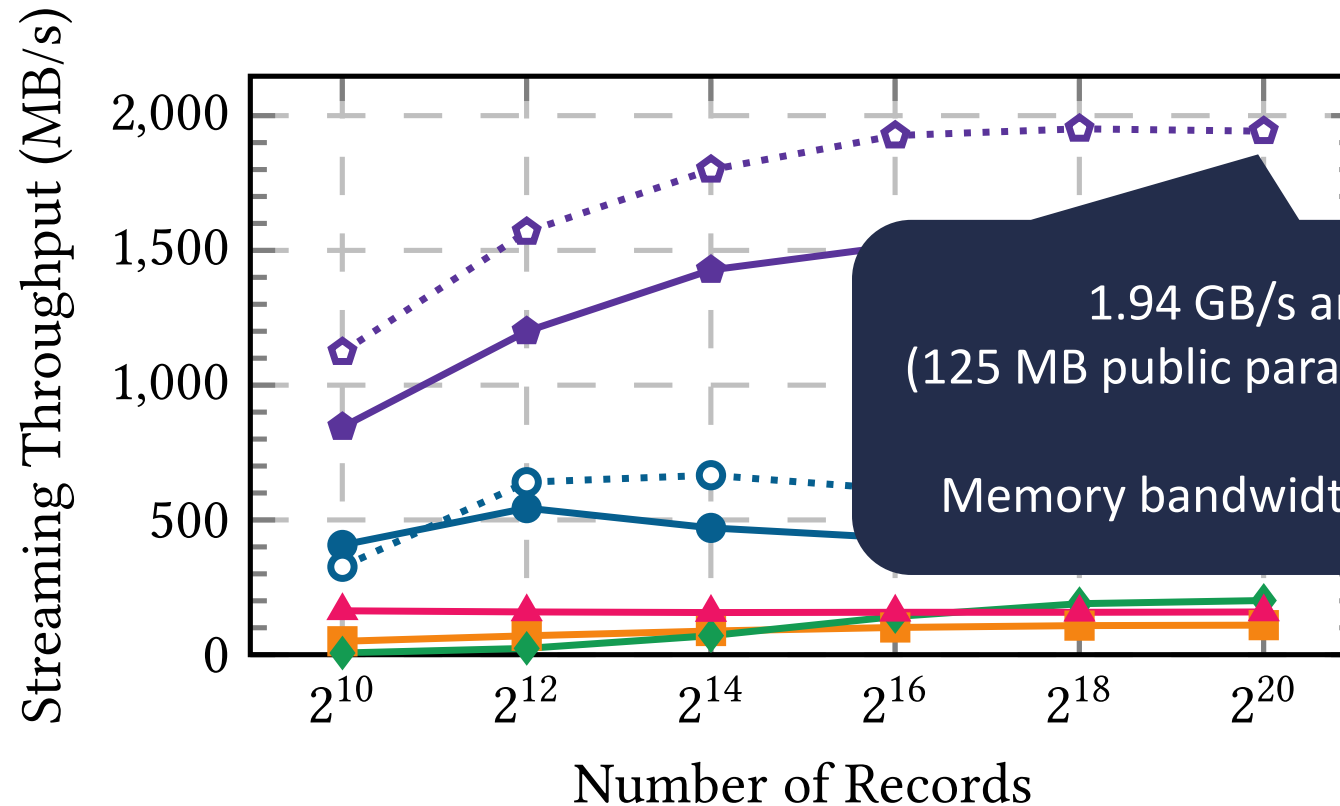
Peaks at ≈ 1.9 GB/s (over 12 \times faster than earlier constructions)

Packed versions rely on response compression (larger public parameters, higher throughput)



The Streaming Setting

Streaming throughput: ignoring query expansion costs, assuming optimal record size for each system

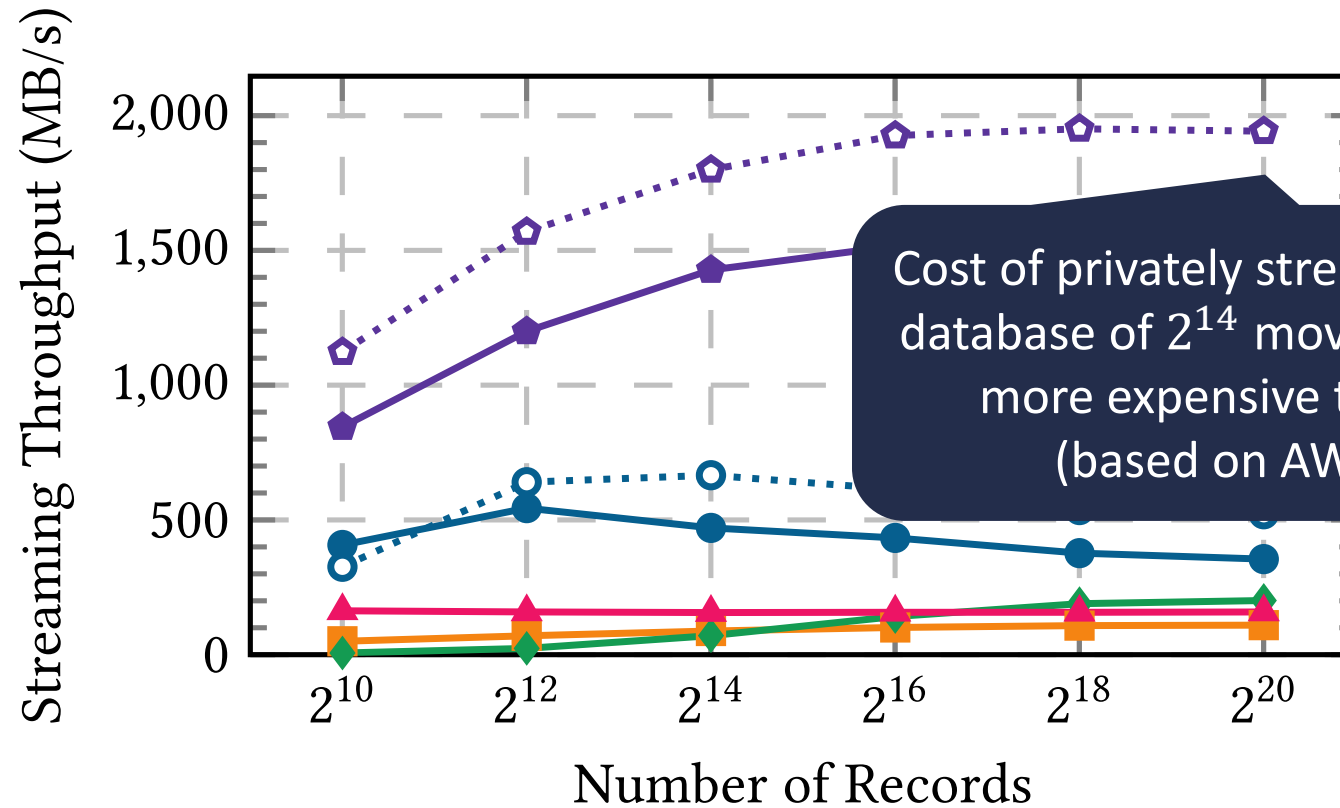


Packing outperforms non-packed protocol for streaming settings



The Streaming Setting

Streaming throughput: ignoring query expansion costs, assuming optimal record size for each system



Cost of privately streaming a 2 GB movie from database of 2^{14} movies estimated to be 1.9× more expensive than direct download (based on AWS compute costs)

Packing outperforms non-packed protocol for streaming settings

- SPIRAL
- SPIRALPACK
- ◆ SPIRALSTREAM
- ◇ SPIRALSTREAMPACK
- SealPIR
- ◆ FastPIR
- ▲ OnionPIR

The SPIRAL Family of PIR

Techniques to translate between FHE schemes enables new trade-offs in single-server PIR

Used for both query compression and response compression

Automatic parameter selection to choose lattice parameters based on database configuration

Base version of SPIRAL

Query size:	14 KB	4.5× smaller
Rate:	0.41	2.1× higher
Throughput:	333 MB/s	2.9× higher

Streaming versions of SPIRAL

Rate:	0.81	3.4× smaller
Throughput:	1.9 GB/s	12.3× higher

(Database with 2^{14} records of size 100 KB)

RESPIRE: The Small Record Setting

Suppose database has **small** records (~ 256 bytes)

Query size: 16 KB

Response size: 20 KB

Throughput: 200 MB/s

Both queries and responses are
much larger than the record!

Reason: LWE ciphertexts are **big**

Recall that query consists of (packed) Regev ciphertext (at least one element of R_q)

- $R = \mathbb{Z}[x]/(x^d + 1)$
- For correctness + security, need $d \sim 2048$ and $q \sim 2^{56}$
- Single ciphertext already ≥ 14 KB

Can we reduce communication when records are small?

RESPIRE: The Small Record Setting

Suppose database has **small** records (~ 256 bytes)

Query size: 16 KB
Response size: 20 KB
Throughput: 200 MB/s

RESPIRE

Query size:	4.1 KB
Response size:	2.0 KB
Throughput:	204 MB/s

3.9× smaller
10× smaller

Reason: LWE ciphertexts are **big**

Recall that query consists of (packed) Regev ciphertext (at least one element of R_q)

- $R = \mathbb{Z}[x]/(x^d + 1)$
- For correctness + security, need $d \sim 2048$ and $q \sim 2^{56}$
- Single ciphertext already ≥ 14 KB

Can we reduce communication when records are small?

Query Expansion, Revisited

Query contains 2^{v_1} matrix Regev encodings



Indicator for index along first dimension

Query contains v_2 GSW encodings



Indicator for index along subsequent dimensions



When database is small, we only need to pack a **small** number of coefficients into an encoding

Each plaintext value is a polynomial of degree d and can hold d values in \mathbb{Z}_q

$$1 + x + x^3 \quad \boxed{1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0} \quad d = 8$$

Query Expansion, Revisited

Query contains 2^{v_1} matrix Regev encodings



Indicator for index along first dimension

Query contains v_2 GSW encodings

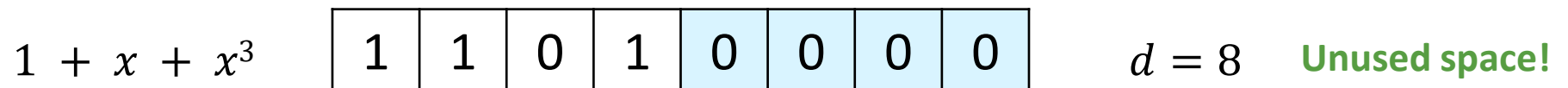


Indicator for index along subsequent dimensions



When database is small, we only need to pack a **small** number of coefficients into an encoding

Each plaintext value is a polynomial of degree d and can hold d values in \mathbb{Z}_q

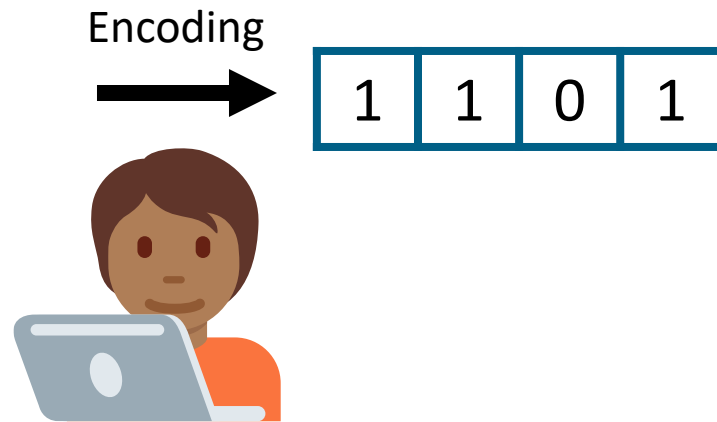


RESPIRE Query Compression

Let d be the ring dimension

If we want to encode (i.e., pack h independent values into a single ciphertext), it suffices to communicate a vector of dimension h rather than d

$$h = 4$$
$$d = 8$$



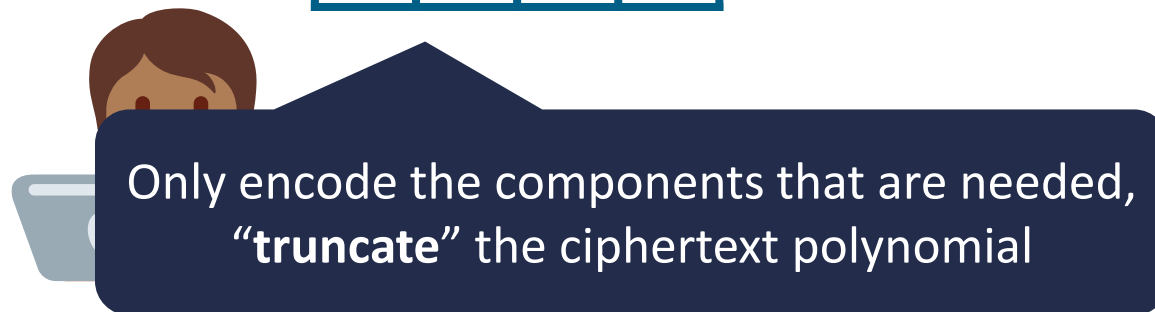
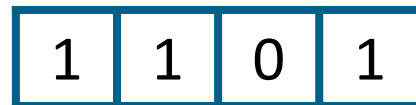
RESPIRE Query Compression

Let d be the ring dimension

If we want to encode (i.e., pack h independent values into a single ciphertext), it suffices to communicate a vector of dimension h rather than d

$$h = 4$$
$$d = 8$$

Encoding

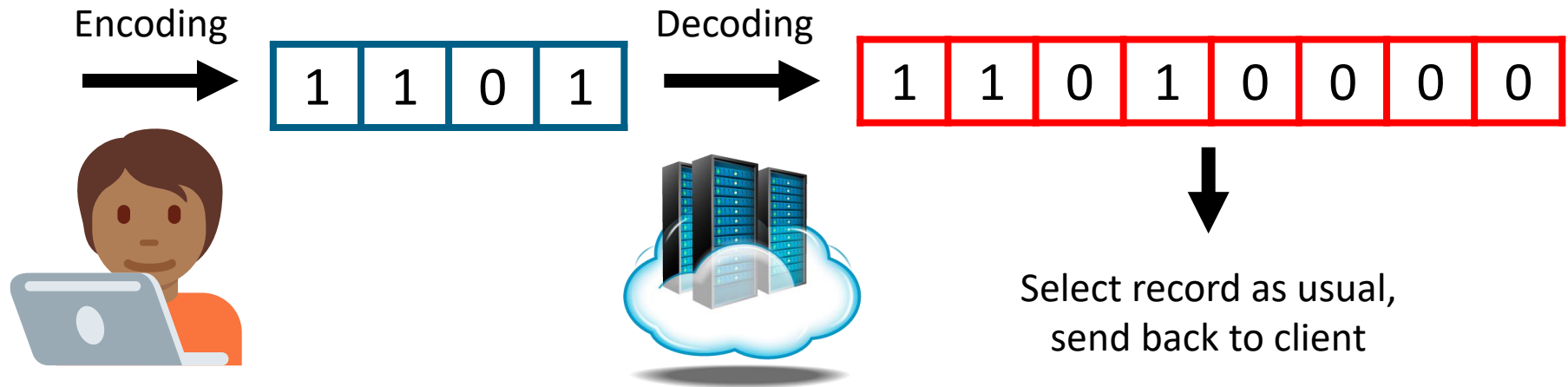


RESPIRE Query Compression

Let d be the ring dimension

If we want to encode (i.e., pack h independent values into a single ciphertext), it suffices to communicate a vector of dimension h rather than d

$$h = 4$$
$$d = 8$$

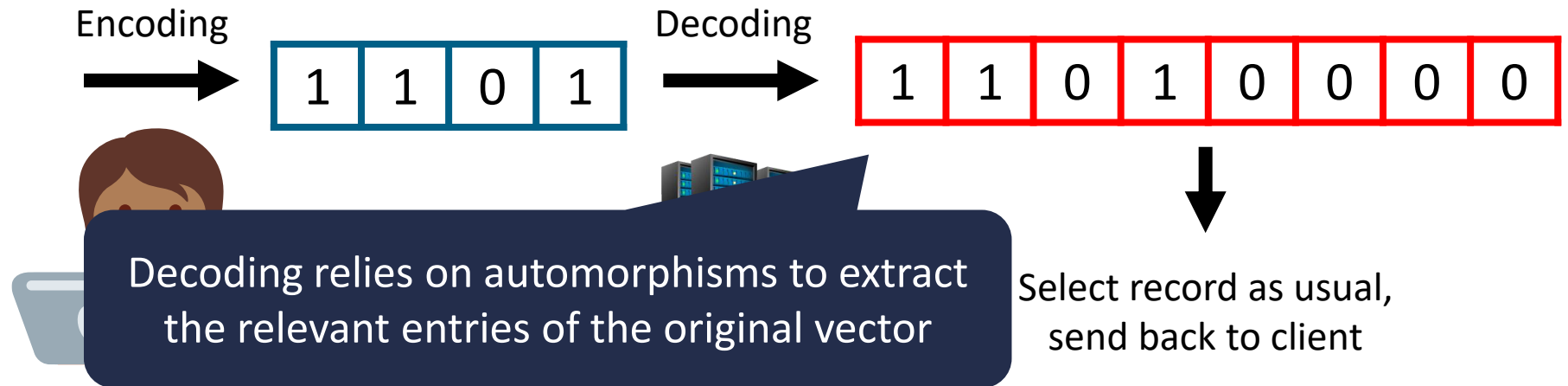


RESPIRE Query Compression

Let d be the ring dimension

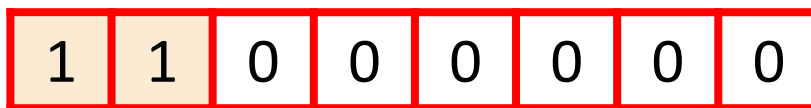
If we want to encode (i.e., pack h independent values into a single ciphertext), it suffices to communicate a vector of dimension h rather than d

$$\begin{aligned} h &= 4 \\ d &= 8 \end{aligned}$$

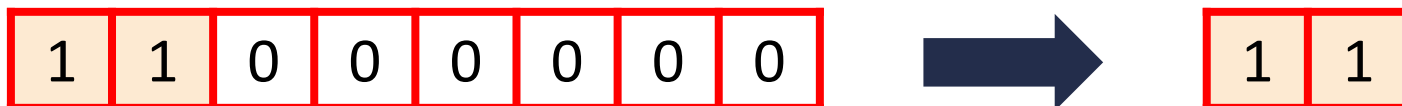


RESPIRE Response Compression

Let d be the ring dimension



Suppose record is *much* smaller than a single ring element



“Ring switching” [BGV12, GHPS12]: translate ciphertext over big ring to a ciphertext over a *subring*

RESPIRE

Query size:	4.1 KB
Response size:	2.0 KB
Throughput:	204 MB/s

Both query and response is “smaller” than standard RLWE ciphertext!

(1 million 256 byte records)

More Recent Developments in PIR

Server preprocessing (client downloads hint at beginning of protocol)

SimplePIR, DoublePIR [HHCMV23]

Very high throughput (nearly memory bandwidth!)

Suitable for databases with small records (a few bits), but has a large hint (hundred of MB)

HintlessPIR [LMRS24], **YPIR** [MW24]

SimplePIR without the hint (by leveraging bootstrapping/key-switching)

Comparable throughput (for big databases), slightly more communication

Piano [ZPSZ23]

Sublinear server computational costs (can scale better to databases that are >100 GB)

Preprocessing phase requires *streaming* the entire database

More Recent Developments in PIR

Server preprocessing (*without* hint)

Doubly-efficient PIR [LMW23]

Server encodes the database to answer queries in sublinear time

Concrete efficiency not yet clear

Many other directions!

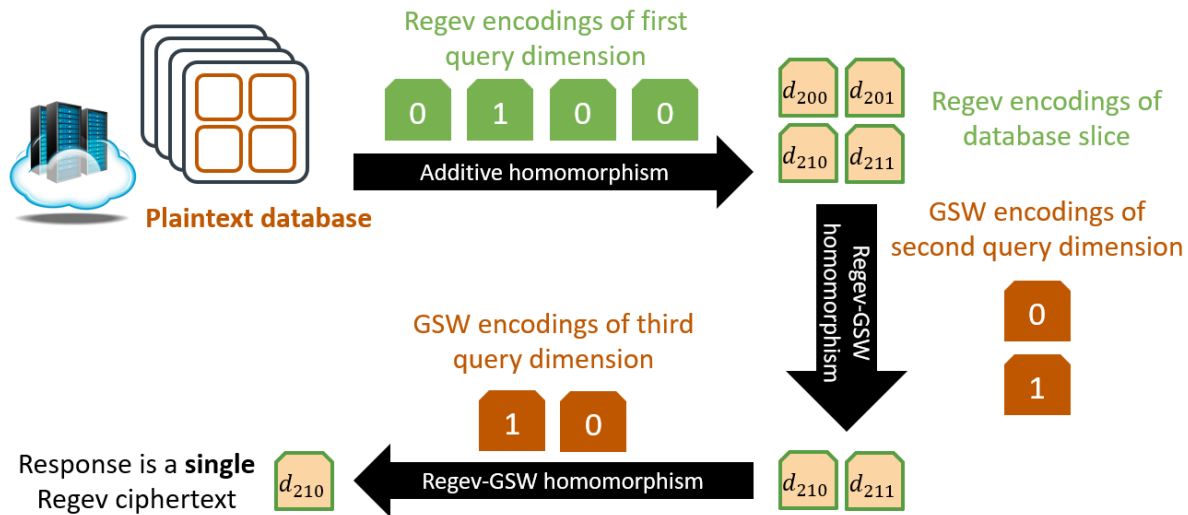
Protocols for batch queries [MR23]

Supporting keyword search [PSY23]

Authenticating the response [CNCWF23]

Takeaway: PIR is an exciting area to work in with many different trade-offs to explore!

SPIRAL and RESPIRE



Techniques to translate between FHE schemes enables new trade-offs in single-server PIR

Useful for both query compression and response compression

SPIRAL: <https://eprint.iacr.org/2022/368.pdf>

RESPIRE: <https://eprint.iacr.org/2024/1165.pdf>

Code: <https://github.com/menonsamir/spiral-rs>

Thank you!