Cerebro: A Platform for Cryptographic Collaborative Learning

Ryan Deng

Joint work with Wenting Zheng, Weikeng Chen,
Aurojit Panda, Raluca Ada Popa, Ion Stoica

CMU, UC Berkeley, NYU

Anti Money Laundering (AML)

Anti Money Laundering (AML)

- Bank wants to detect and prevent money laundering using machine learning.

Anti Money Laundering (AML)

- Bank wants to detect and prevent money laundering using machine learning.
  - Scotiabank CRO: “getting AML right is of **critical strategic importance** to our bank.”

Anti Money Laundering (AML)

- Bank wants to detect and prevent money laundering using machine learning.
  - Scotiabank CRO: “getting AML right is of **critical strategic importance** to our bank.”
- Criminals conceal illegal activities across many banks.

Anti Money Laundering (AML)

- Bank wants to detect and prevent money laundering using machine learning.
  - Scotiabank CRO: “getting AML right is of **critical strategic importance** to our bank.”
- Criminals conceal illegal activities across many banks.

Anti Money Laundering (AML)

- Bank wants to detect and prevent money laundering using machine learning.
  - Scotiabank CRO: “getting AML right is of **critical strategic importance** to our bank.”
- Criminals conceal illegal activities across many banks.

Anti Money Laundering

Anti Money Laundering

- Want to jointly compute a model for detecting money laundering across many banks.

Anti Money Laundering

- Want to jointly compute a model for detecting money laundering across many banks.
  - Scotiabank CRO: “So in the future, **collaboration** will be **vital** … The ability to put together our data sets and **collaborate** … will enhance yields by **orders of magnitude**.”

Anti Money Laundering

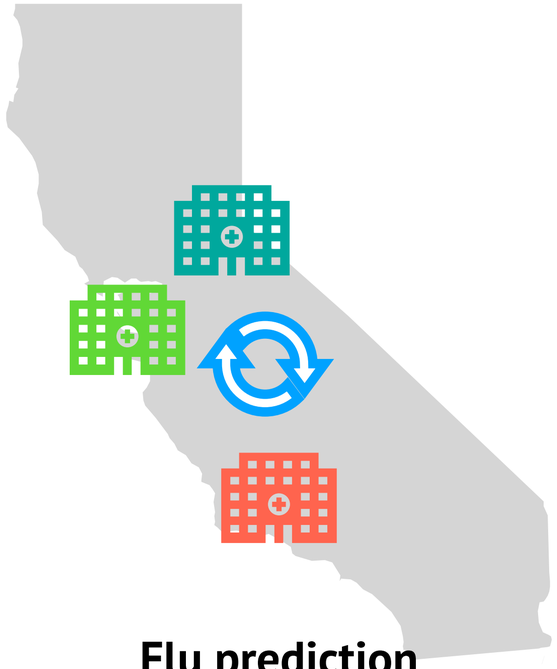
- Want to jointly compute a model for detecting money laundering across many banks.
  - Scotiabank CRO: “So in the future, **collaboration** will be **vital** … The ability to put together our data sets and **collaborate** … will enhance yields by **orders of magnitude**.”

Anti Money Laundering

- Want to jointly compute a model for detecting money laundering across many banks.
  - Scotiabank CRO: “So in the future, **collaboration** will be **vital** … The ability to put together our data sets and **collaborate** … will enhance yields by **orders of magnitude**.”
- **Cannot share data due to privacy concerns, regulatory policies, and business competition.**

Anti Money-Laundering

Anti Money-Laundering



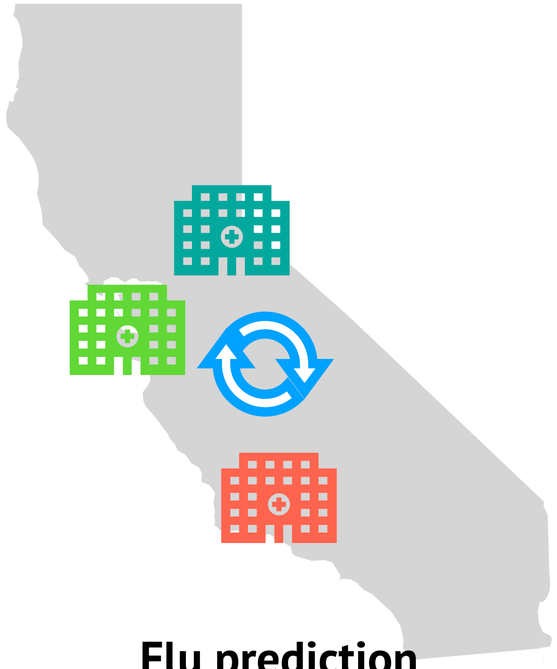
Flu prediction

Anti Money-Laundering

Flu prediction

Cloud-based video analysis for home intrusion

Anti Money-Laundering

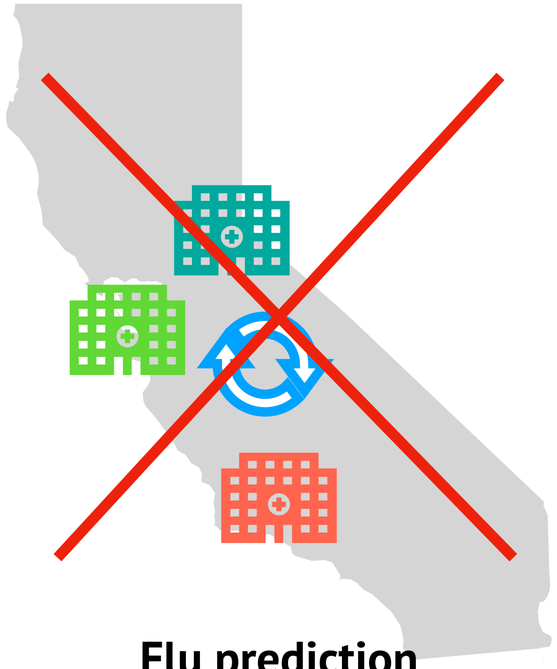


Flu prediction

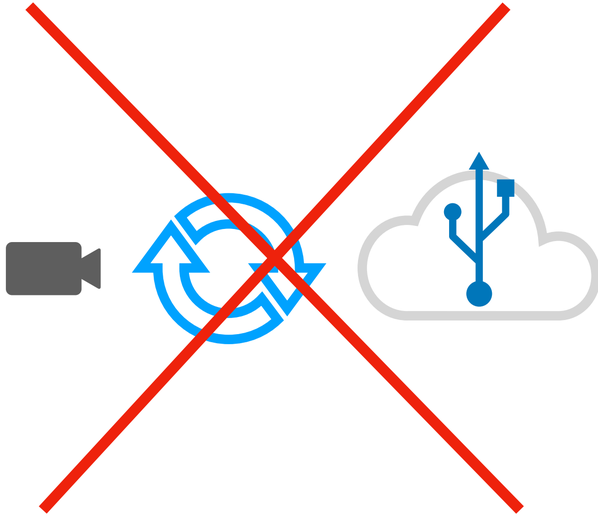
Cloud-based video analysis for home intrusion

...

Anti Money-Laundering



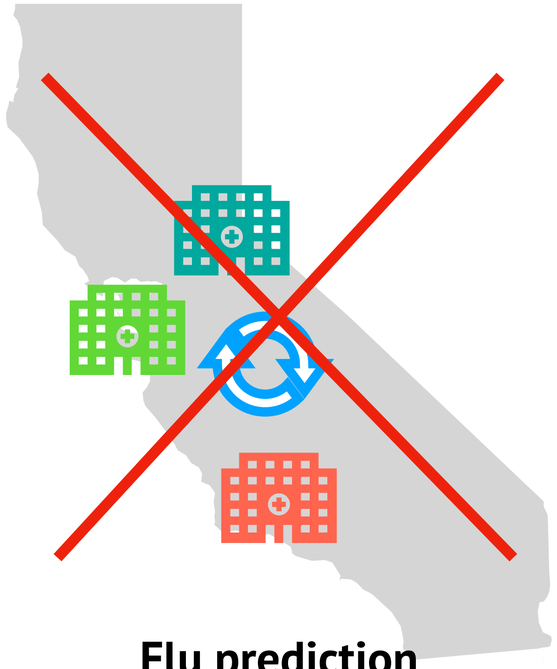
Flu prediction



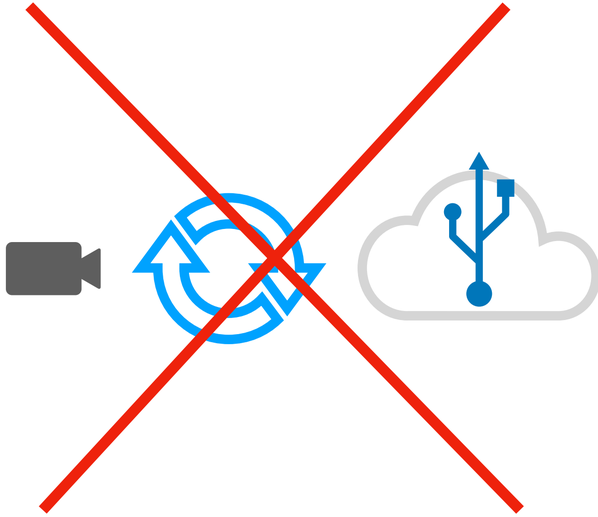
Cloud-based video analysis for home intrusion

...

Anti Money-Laundering



Flu prediction



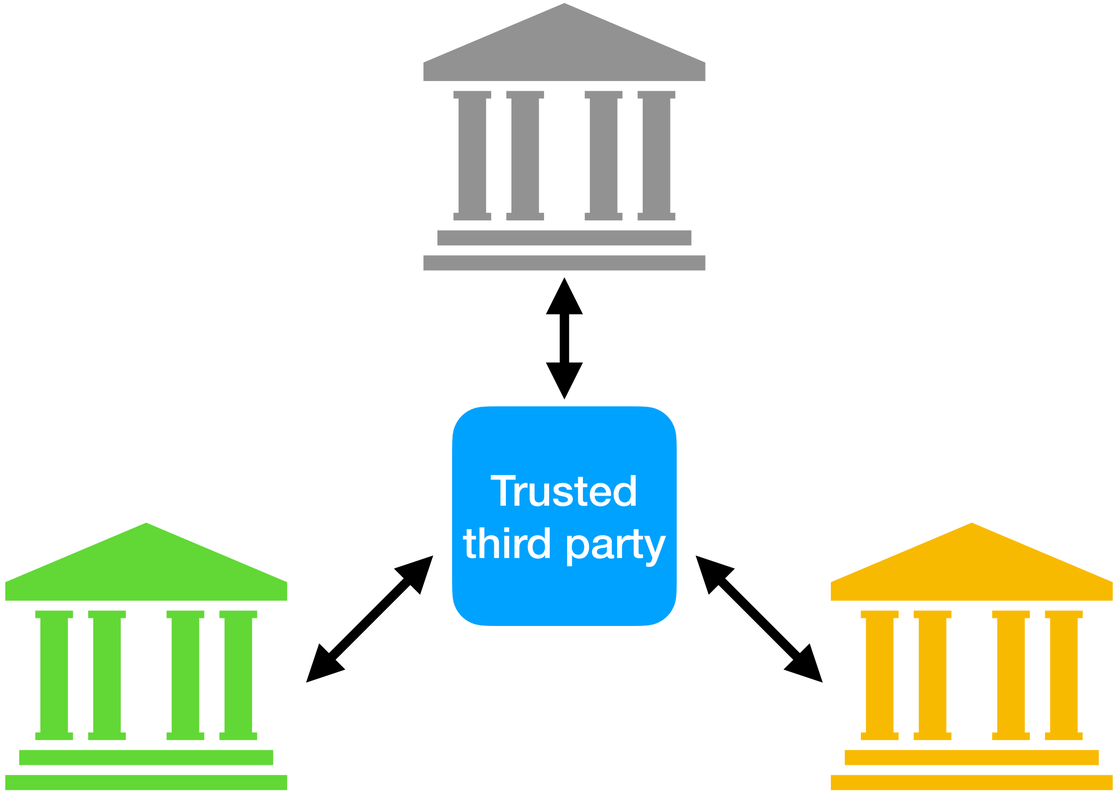
Cloud-based video analysis for home intrusion

...

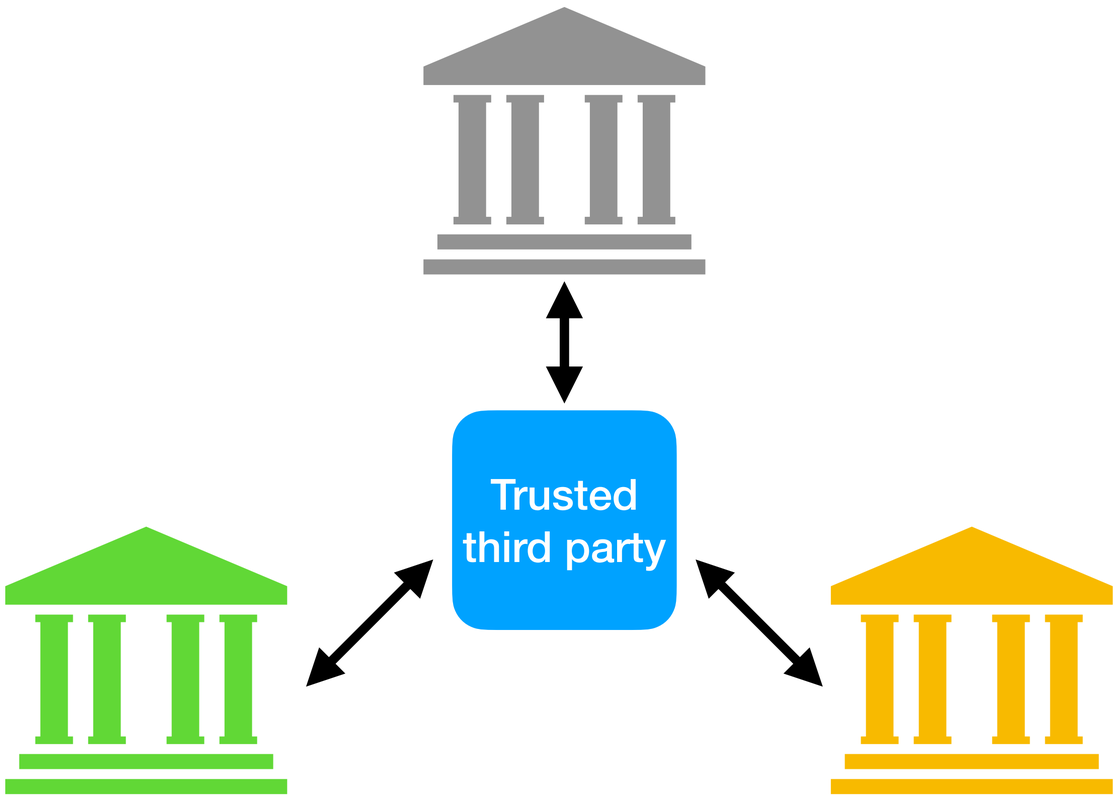
People are unwilling or unable to share sensitive data for collaborative computation

How to allow organizations to
share data for collaboration
without showing the plaintext data?

Secure multiparty computation (MPC[Yao82,GMW87,BGW88])



Secure multiparty computation (MPC[Yao82,GMW87,BGW88])



- Parties emulate a trusted third party via cryptography

Secure multiparty computation (MPC[Yao82,GMW87,BGW88])

- Parties emulate a trusted third party via cryptography

Secure multiparty computation (MPC[Yao82,GMW87,BGW88])

- Parties emulate a trusted third party via cryptography
- No party learns any information beyond the final result

Challenges in Deploying MPC End-to-End

Challenges in Deploying MPC End-to-End

- Generality vs. Performance

Challenges in Deploying MPC End-to-End

- Generality vs. Performance
- Privacy vs. Transparency

Generality vs. Performance

Generality vs. Performance

Option 1: Specialized Solution	

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task.	

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. SecureML [MZ17] Gazelle [JVC18] XONN [RSCLLK19] Spindle [FTPSSBH21] Privacy-preserving ridge regression [NWIJBT13]	

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task.	

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient	

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient ✗ Not generalizable	

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient ✗ Not generalizable	Utilize a generic MPC framework.

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient ✗ Not generalizable	Utilize a generic MPC framework. SPDZ [DPSZ11]

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient ✗ Not generalizable	Utilize a generic MPC framework. SPDZ [DPSZ11] AG-MPC [WKR17]

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient ✗ Not generalizable	Utilize a generic MPC framework. SPDZ [DPSZ11] → MASCOT[KOS16] SPDZ [DPSZ11] → Vectorized[MZ17] SPDZ [DPSZ11] → Overdrive[KPR18]

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient ✗ Not generalizable	Utilize a generic MPC framework.

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task. ✓ Can be very efficient ✗ Not generalizable	Utilize a generic MPC framework. ✓ Generic functionality

Generality vs. Performance

Option 1: Specialized Solution	Option 2: Generic Solution
Use a hand-tuned MPC protocol for a specific learning task.	Utilize a generic MPC framework.
✓ Can be very efficient	✓ Generic functionality
✗ Not generalizable	✗ Difficult to use efficiently

Challenges in Deploying MPC End-to-End

- Generality vs. Performance

Challenges in Deploying MPC End-to-End

- Generality vs. Performance
- Privacy vs. Transparency

Challenges in Deploying MPC End-to-End

- Generality vs. Performance
- Privacy vs. Transparency
  - Policies

Challenges in Deploying MPC End-to-End

- Generality vs. Performance
- Privacy vs. Transparency
  - Policies
  - Auditing

Before Release: Policies

Before Release: Policies

- What if my competitors benefit more than me?

Before Release: Policies

- What if my competitors benefit more than me?

60%

Before Release: Policies

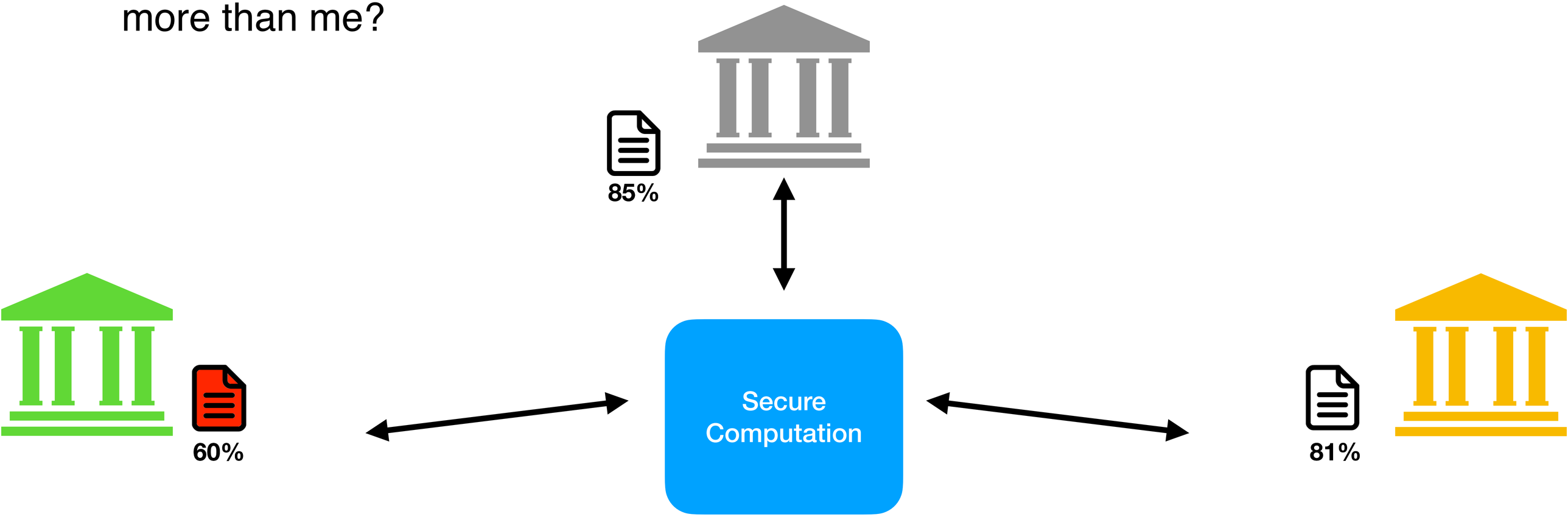
- What if my competitors benefit more than me?

Before Release: Policies

- What if my competitors benefit more than me?

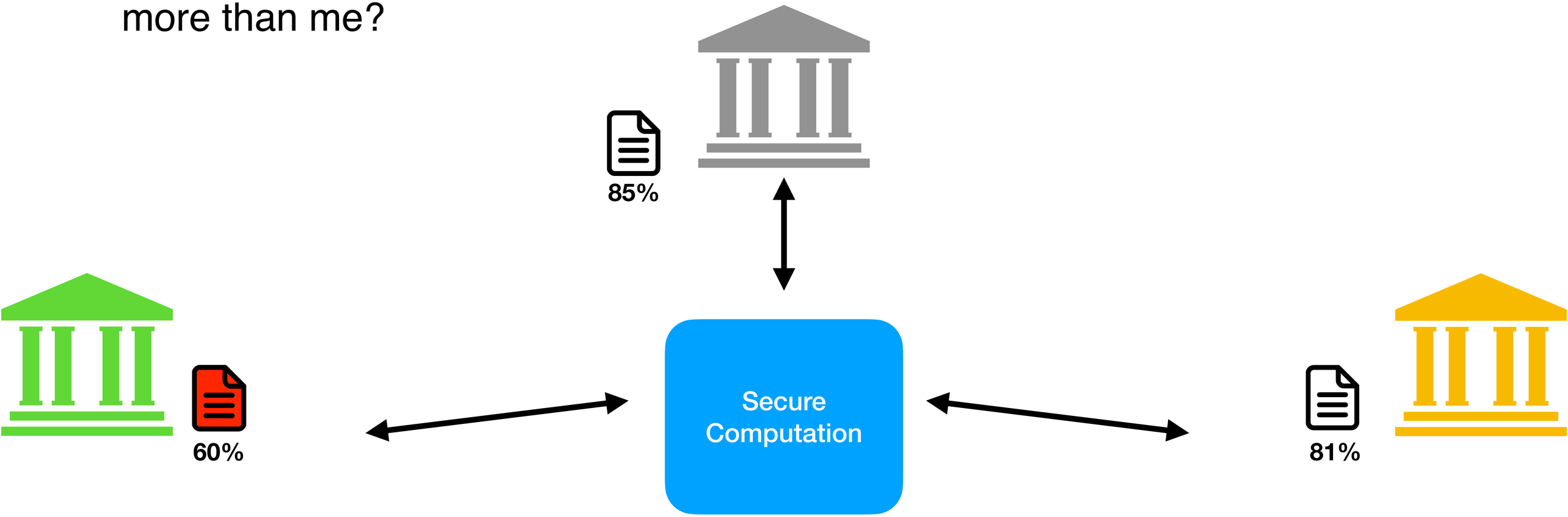
Before Release: Policies

- What if my competitors benefit more than me?



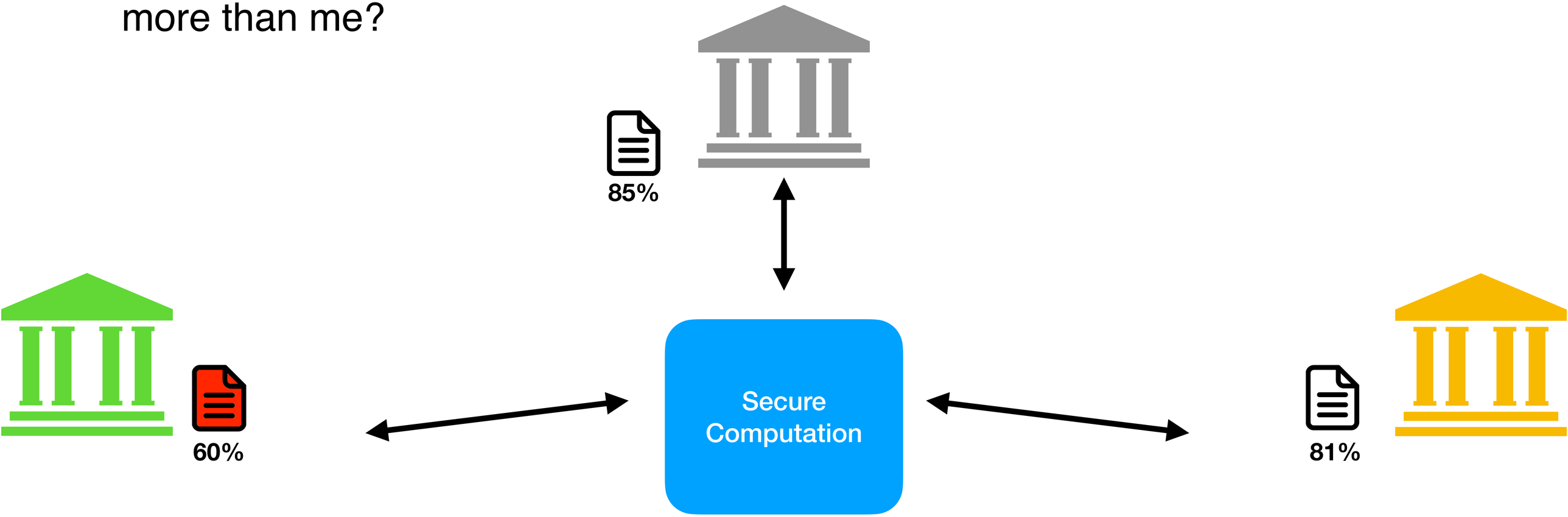
Before Release: Policies

- What if my competitors benefit more than me?



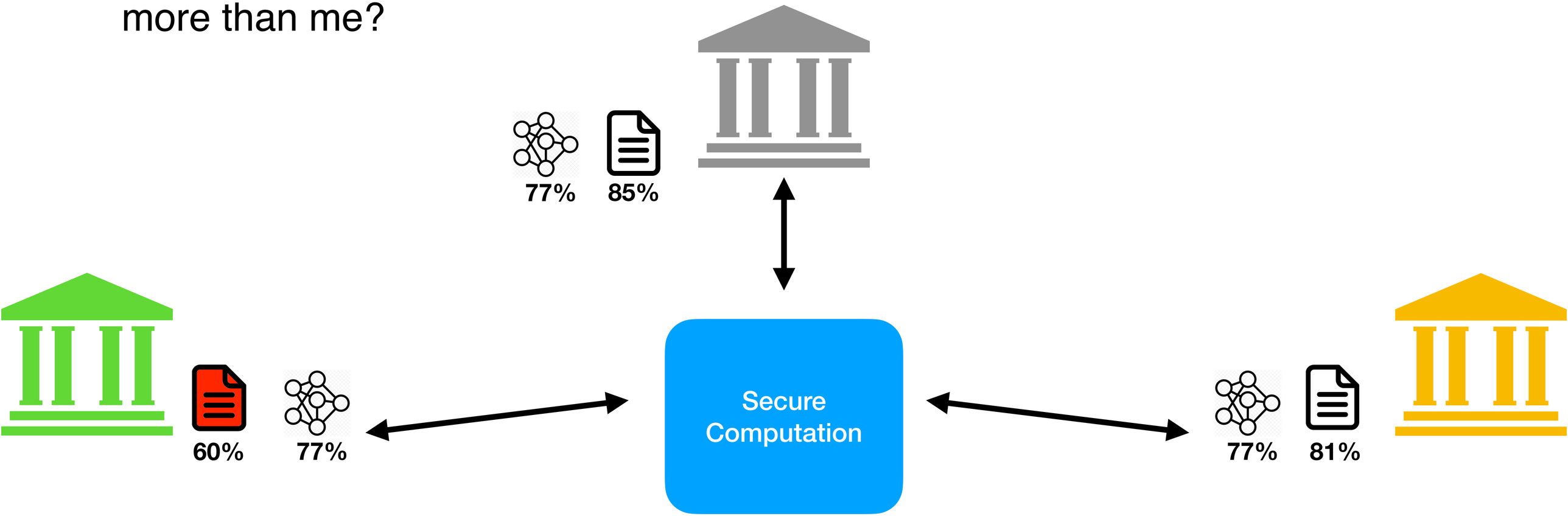
Before Release: Policies

- What if my competitors benefit more than me?



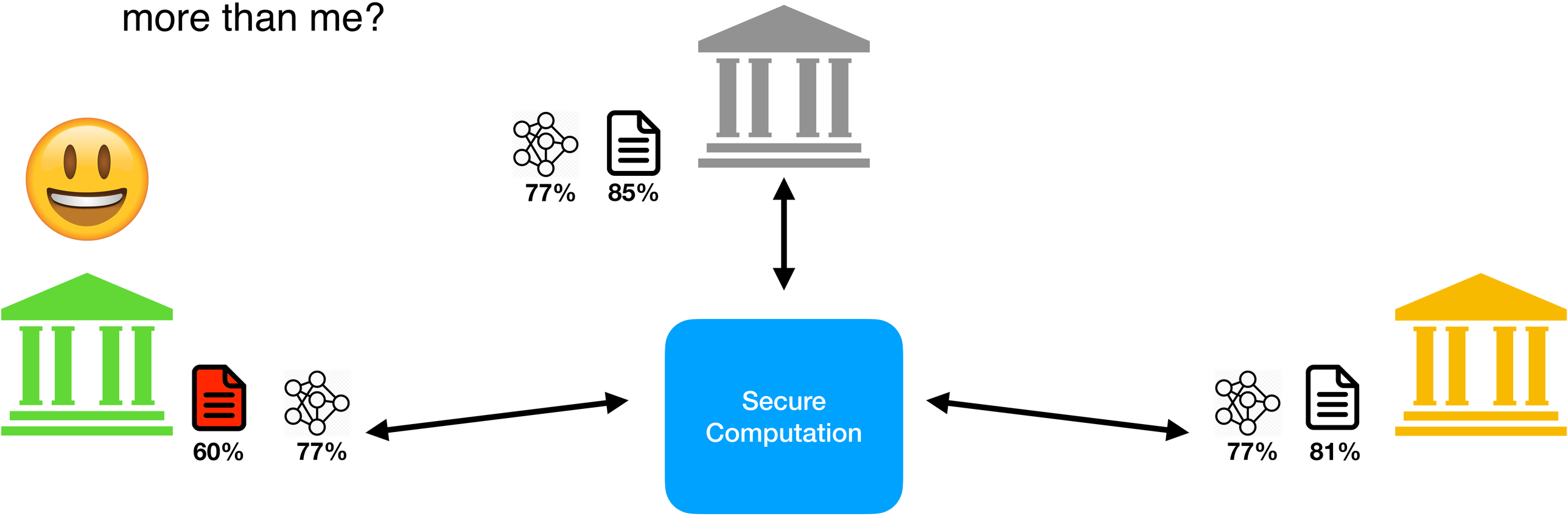
Before Release: Policies

- What if my competitors benefit more than me?



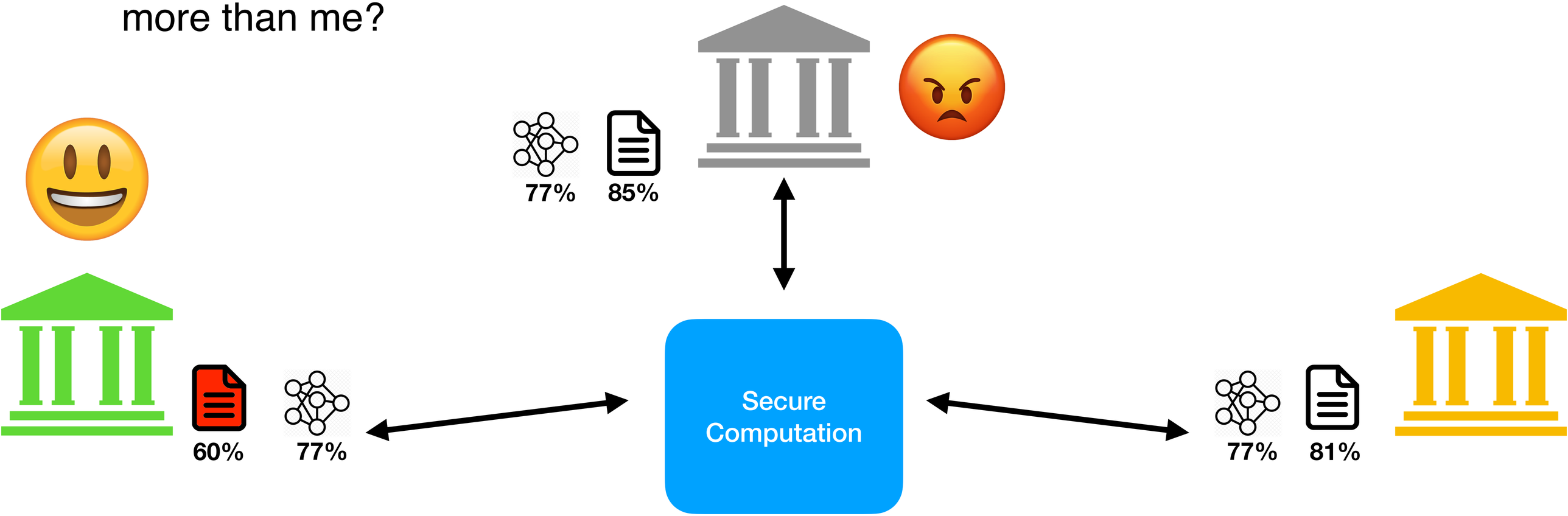
Before Release: Policies

- What if my competitors benefit more than me?



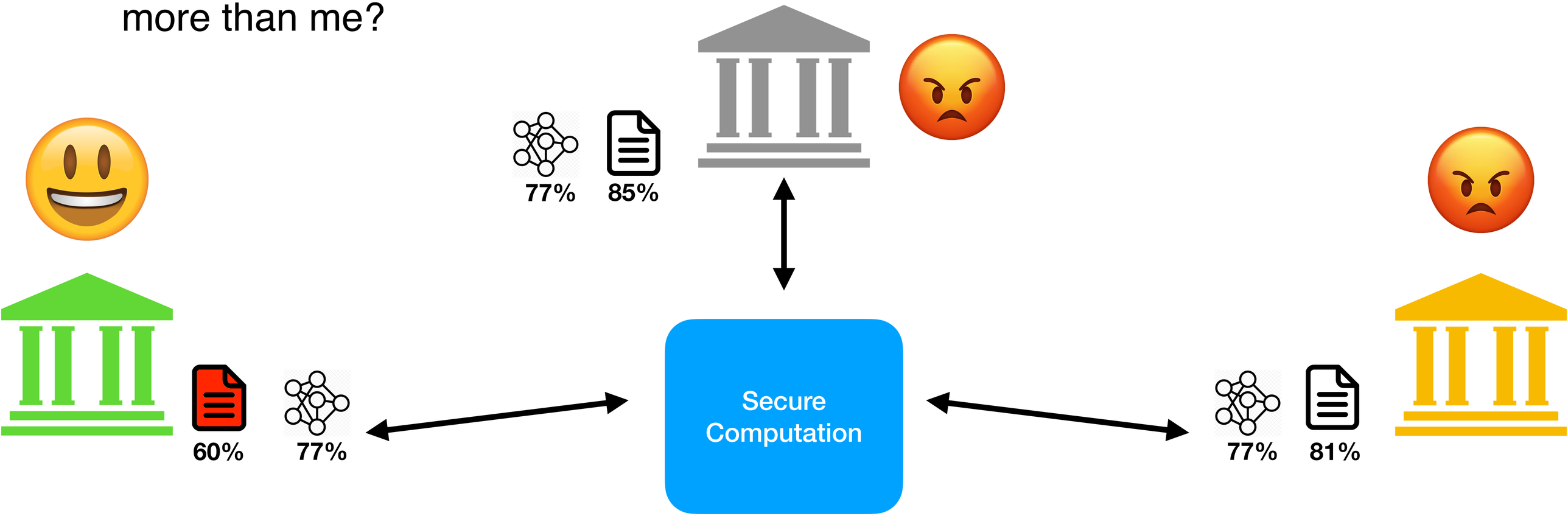
Before Release: Policies

- What if my competitors benefit more than me?



Before Release: Policies

- What if my competitors benefit more than me?



Before Release: Policies

- What if my competitors benefit significantly more than me?

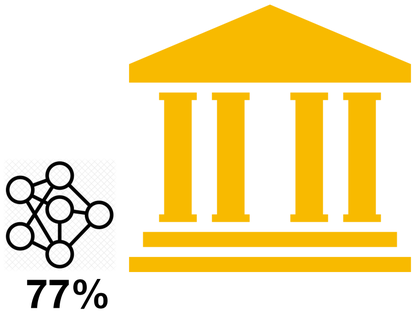
Before Release: Policies

- What if my competitors benefit significantly more than me?
- What happens if the released model leaks information about the underlying training data?

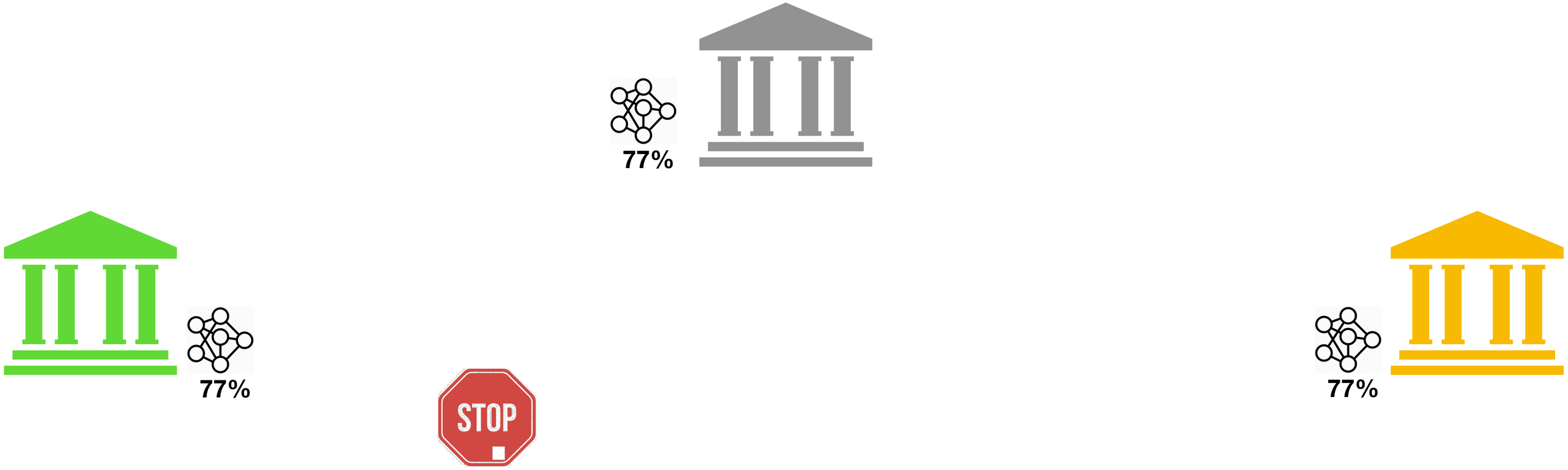
Before Release: Policies

- What if my competitors benefit significantly more than me?
- What happens if the released model leaks information about the underlying training data?
- What happens if the released model does not meet a minimum performance threshold?

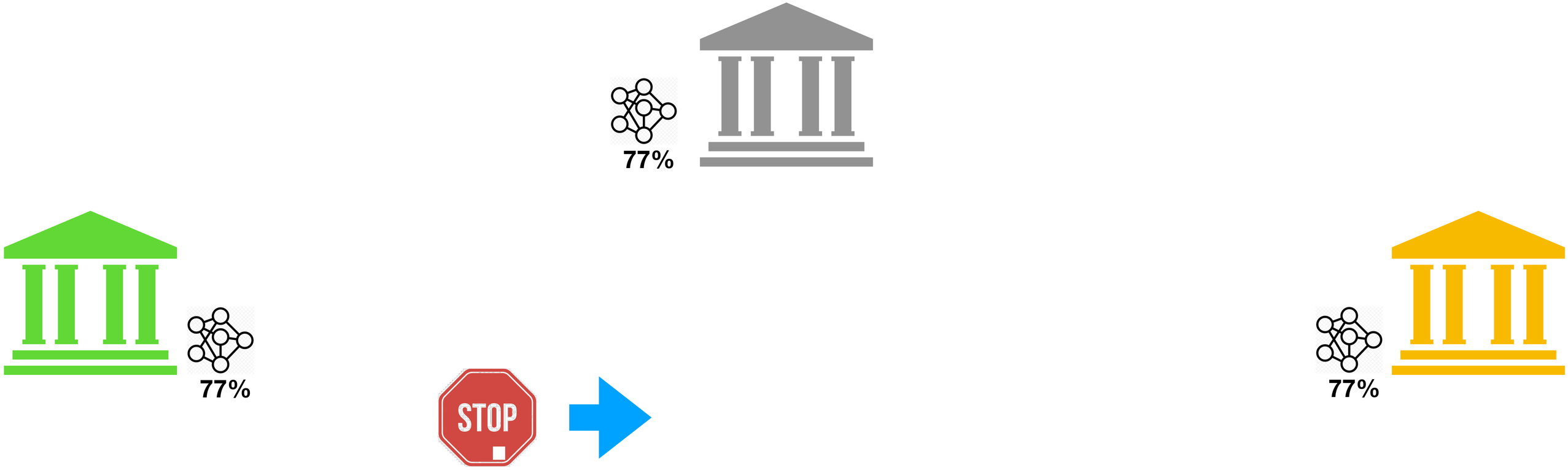
After Release: Auditing



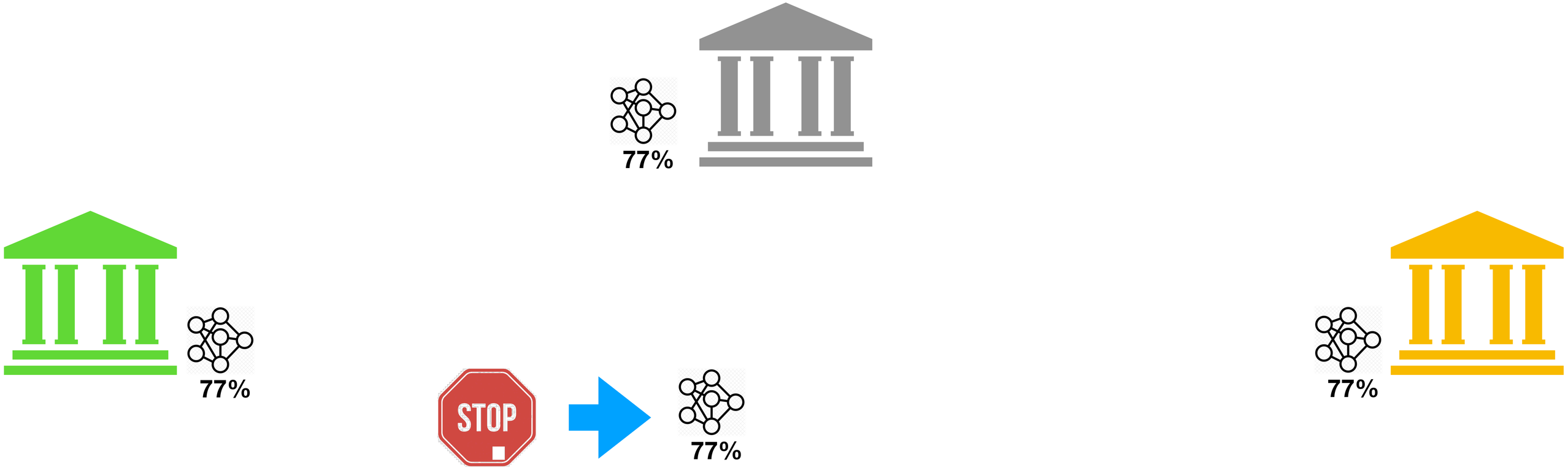
After Release: Auditing



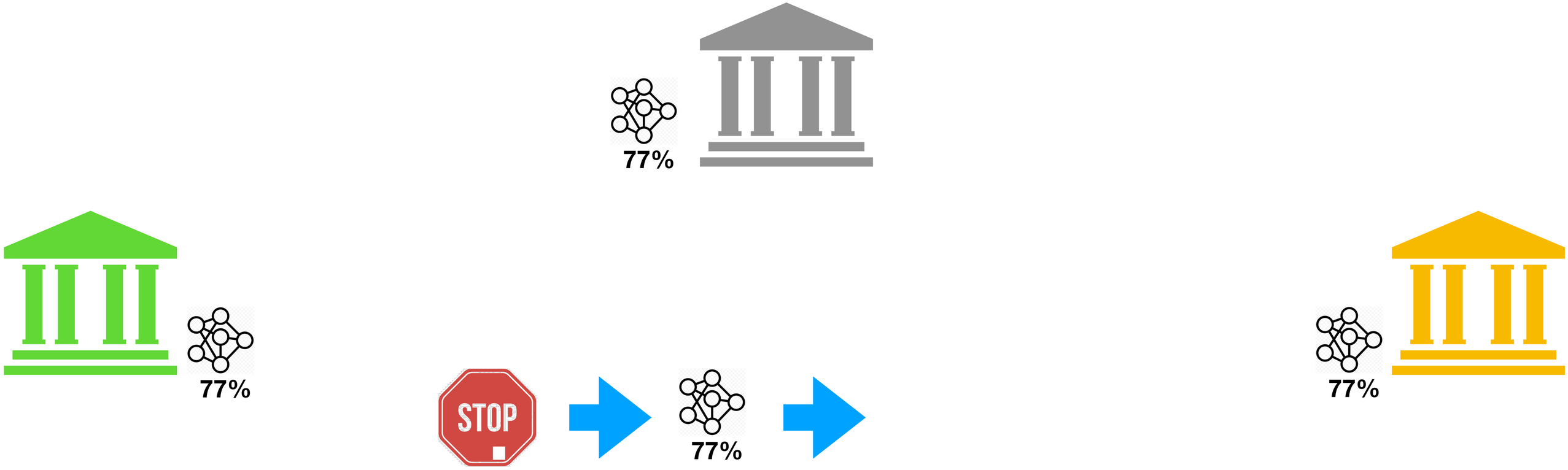
After Release: Auditing



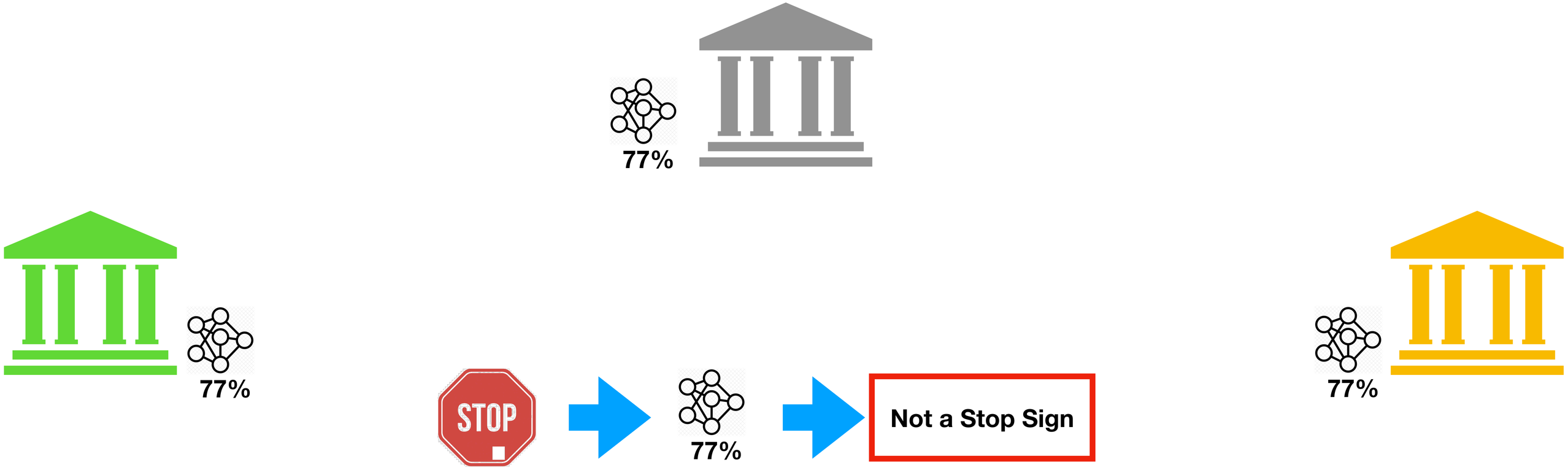
After Release: Auditing



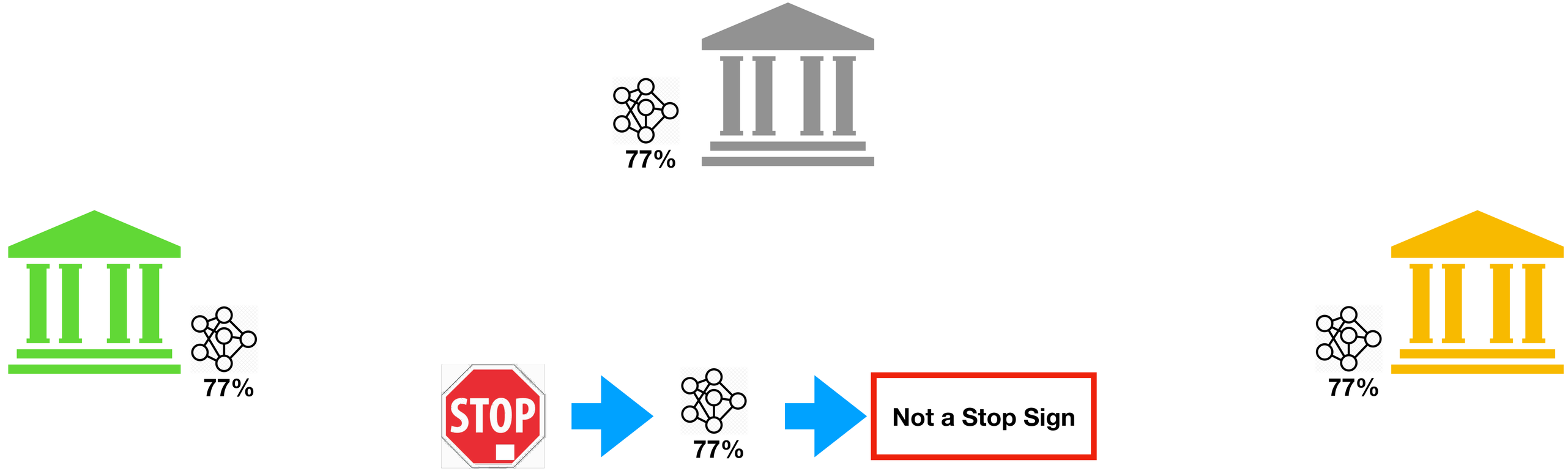
After Release: Auditing



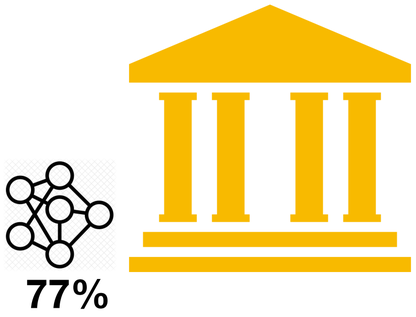
After Release: Auditing



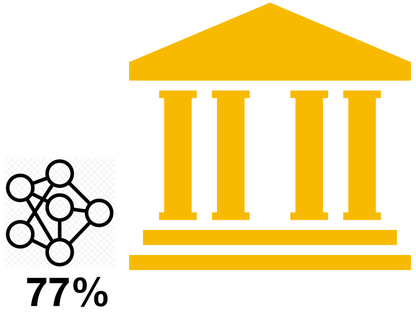
After Release: Auditing



After Release: Auditing



After Release: Auditing



How do we trace bad data to the source?

Cerebro

Cerebro

End-to-end platform for secure collaborative learning

Cerebro

End-to-end platform for secure collaborative learning

Threat Model

Cerebro

End-to-end platform for secure collaborative learning

Threat Model

- N-party dishonest majority.

Cerebro

End-to-end platform for secure collaborative learning

Threat Model

- N-party dishonest majority.
- Protects against semi-honest and malicious adversaries.

Challenges

Challenges

Generality vs. Performance

Challenges

Generality vs. Performance

Privacy vs. Transparency

Cerebro's Techniques

Generality vs. Performance → Domain Specific Language (DSL)

Generality vs. Performance → MPC Compiler

Privacy vs. Transparency → Release Policies

Privacy vs. Transparency → Auditing Framework

Cerebro's Techniques

Generality vs. Performance → Domain Specific Language (DSL)

Generality vs. Performance → **MPC Compiler**

Privacy vs. Transparency → Release Policies

Privacy vs. Transparency → Auditing Framework

Cerebro's Techniques

Generality vs. Performance

- Domain Specific Language (DSL)

  Please see paper

- **MPC Compiler**

Privacy vs. Transparency

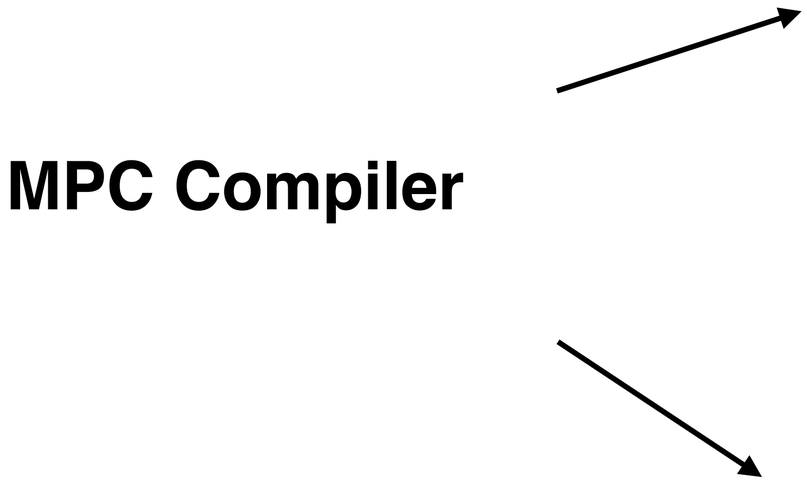
- Release Policies

  Please see paper

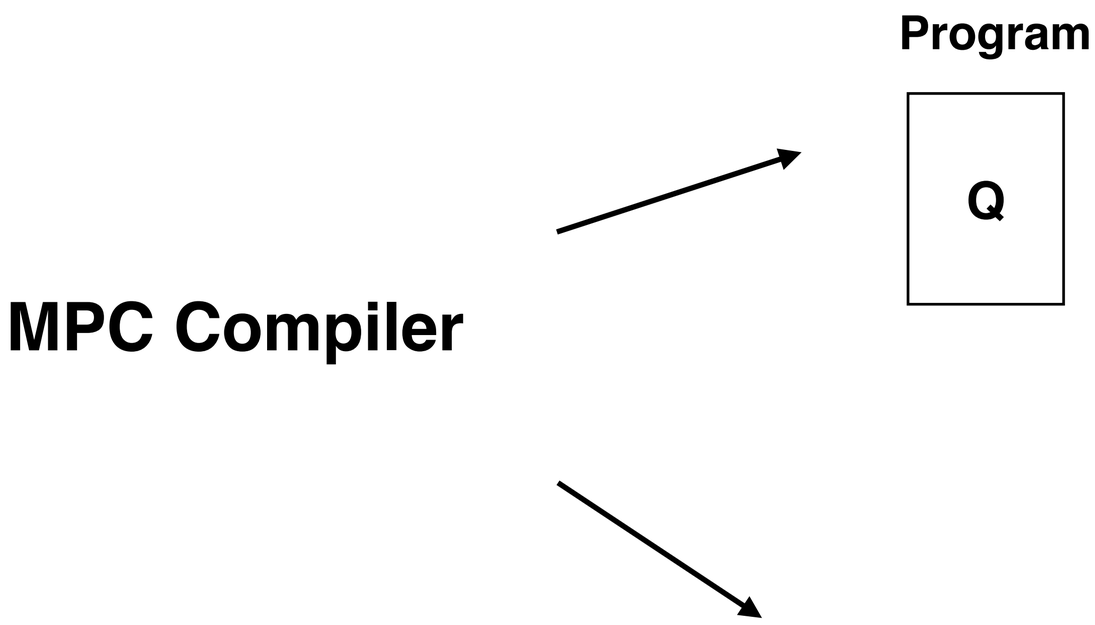
- Auditing Framework

  Please see paper

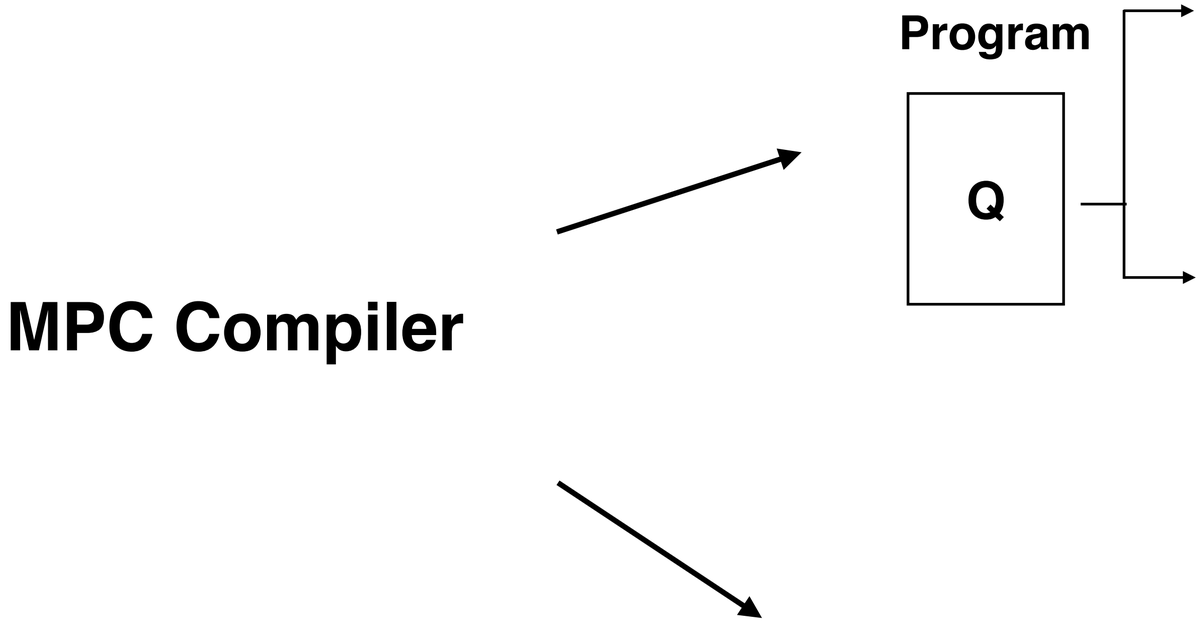
Cerebro’s MPC Compiler



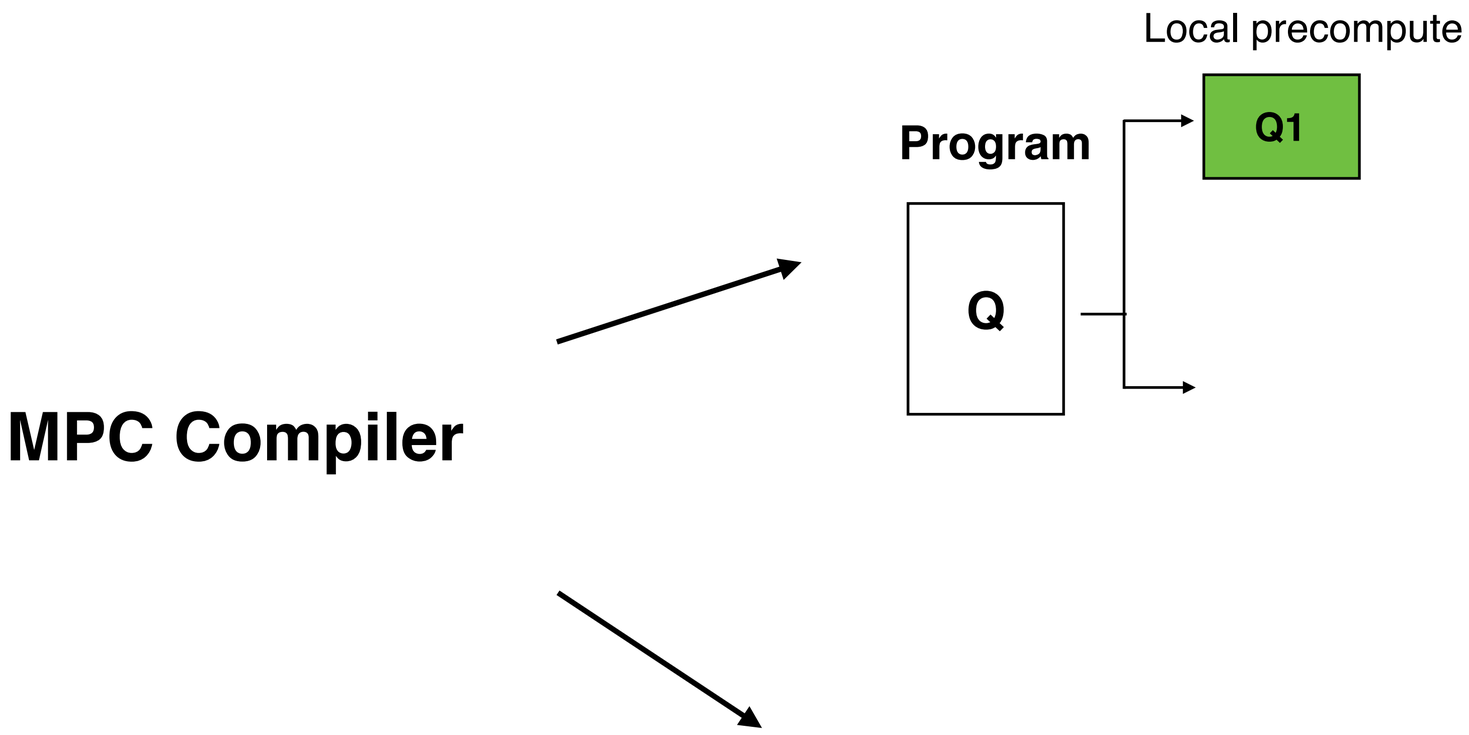
Cerebro's MPC Compiler



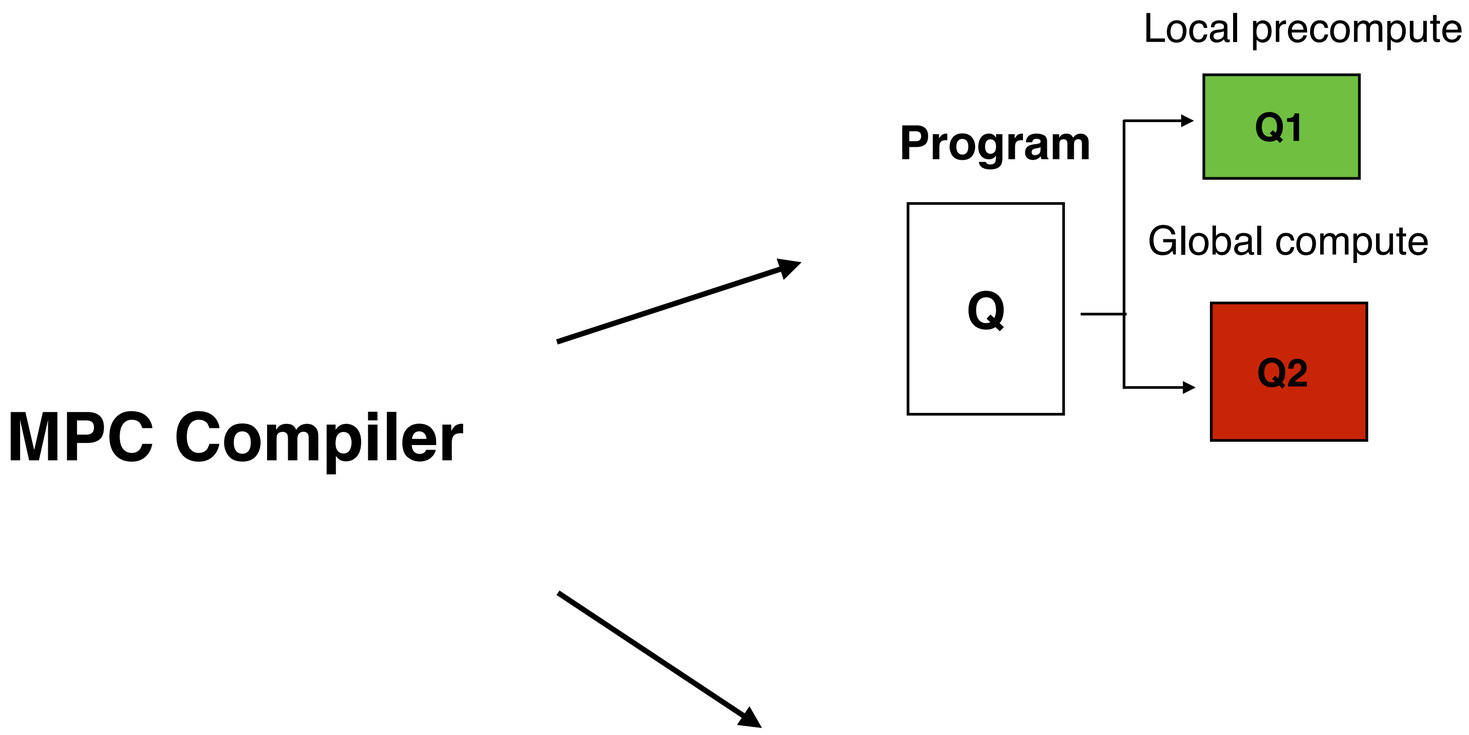
Cerebro's MPC Compiler



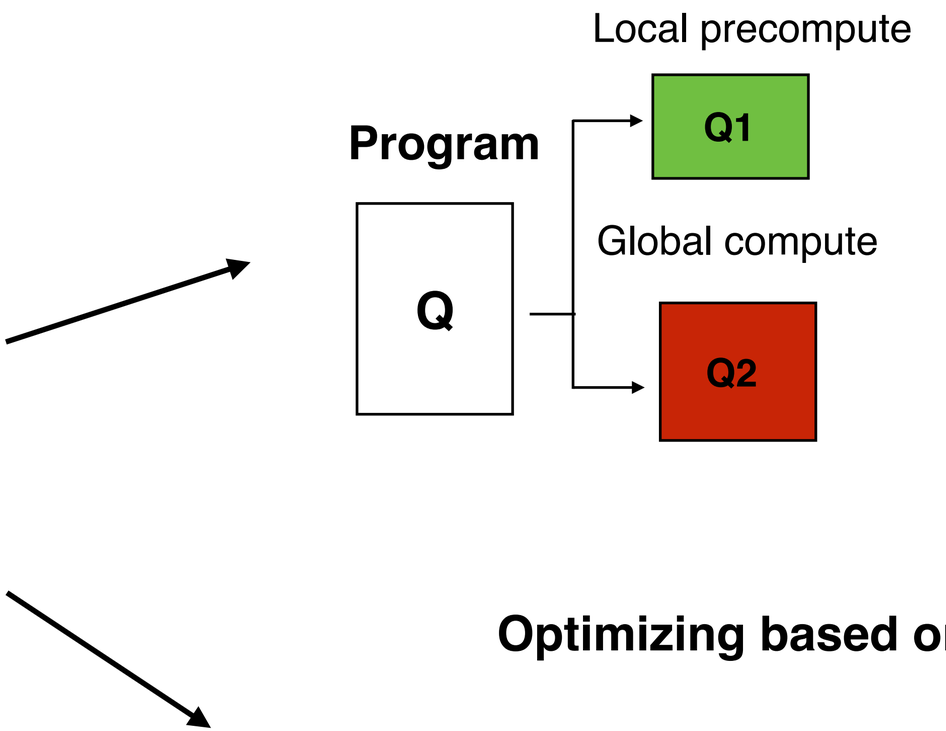
Cerebro's MPC Compiler



Cerebro's MPC Compiler



Cerebro's MPC Compiler

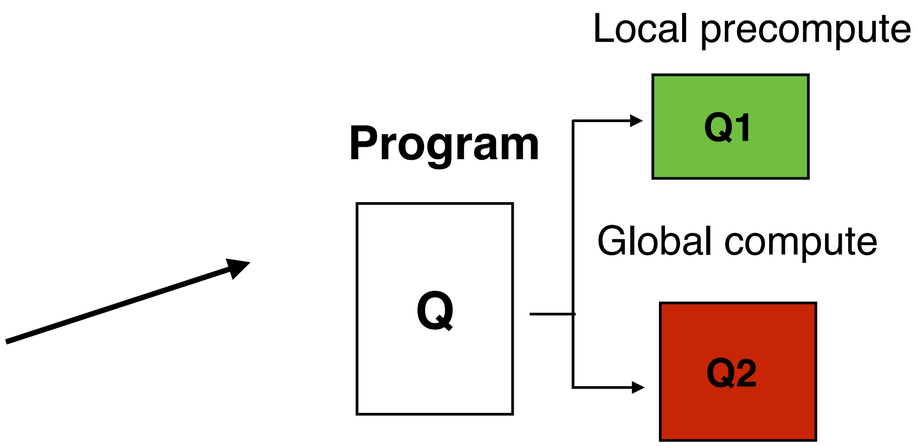


MPC Compiler

Optimizing based on network layout

Cerebro's MPC Compiler

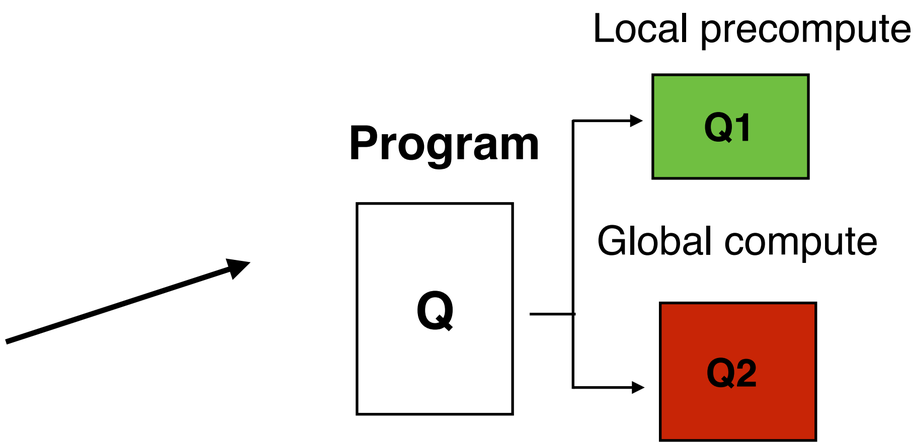
MPC Compiler

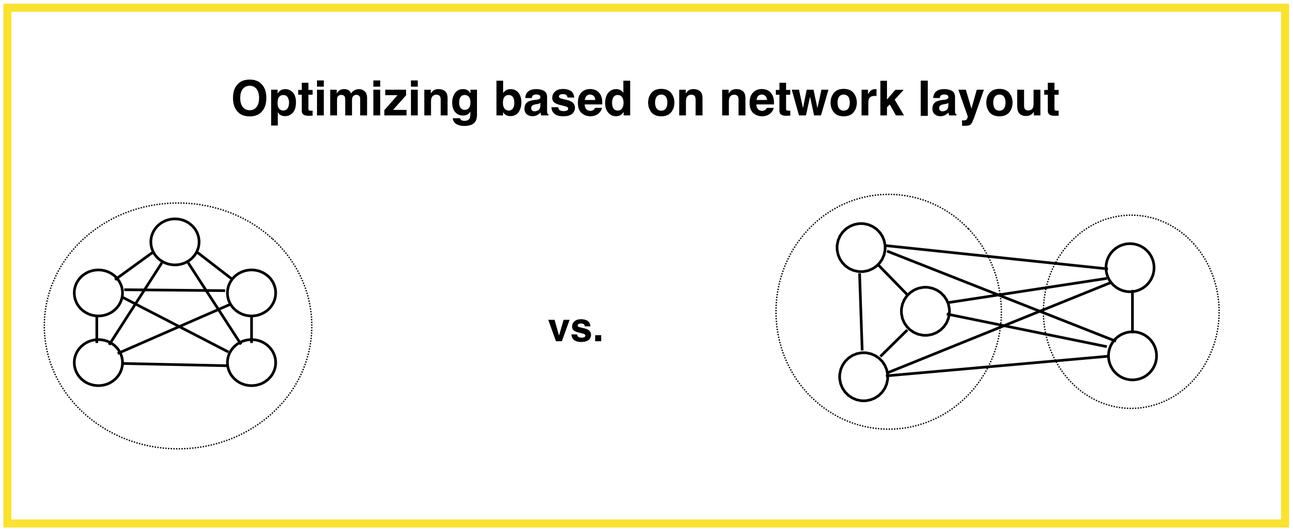


Optimizing based on network layout

Cerebro's MPC Compiler

MPC Compiler



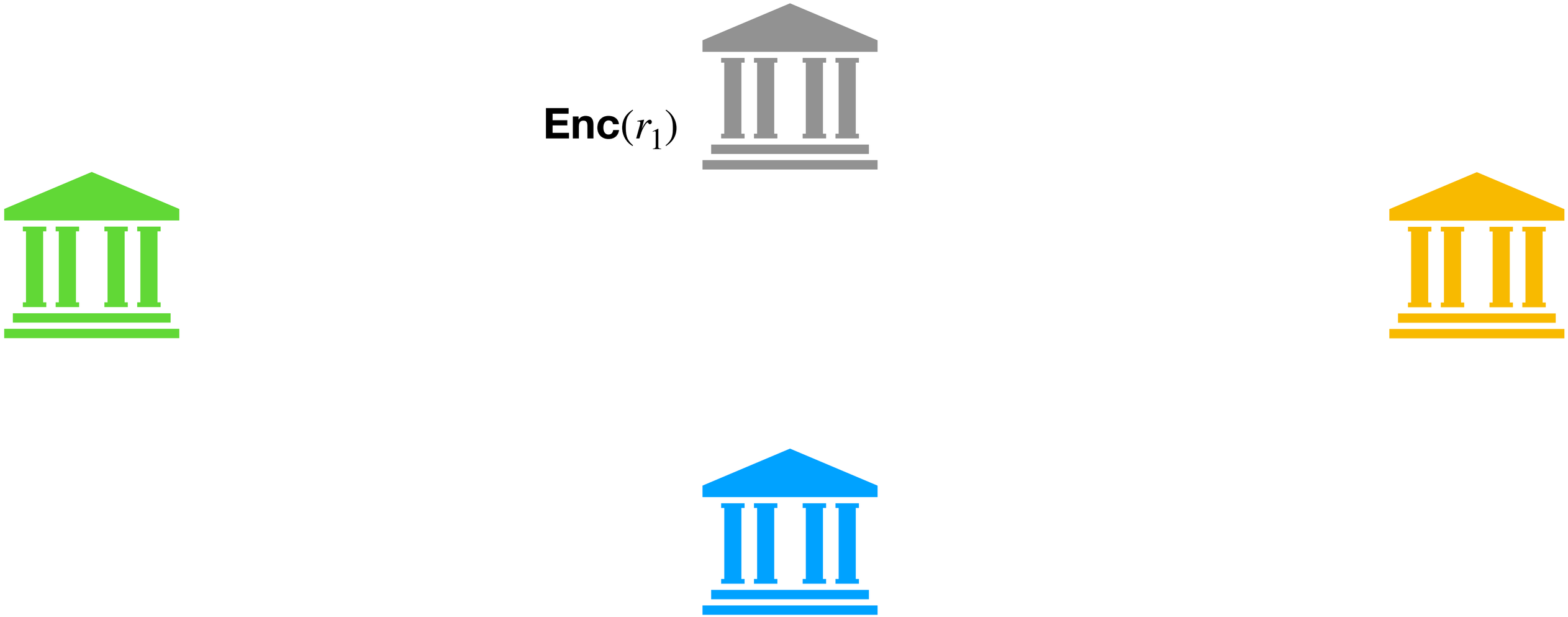


Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]

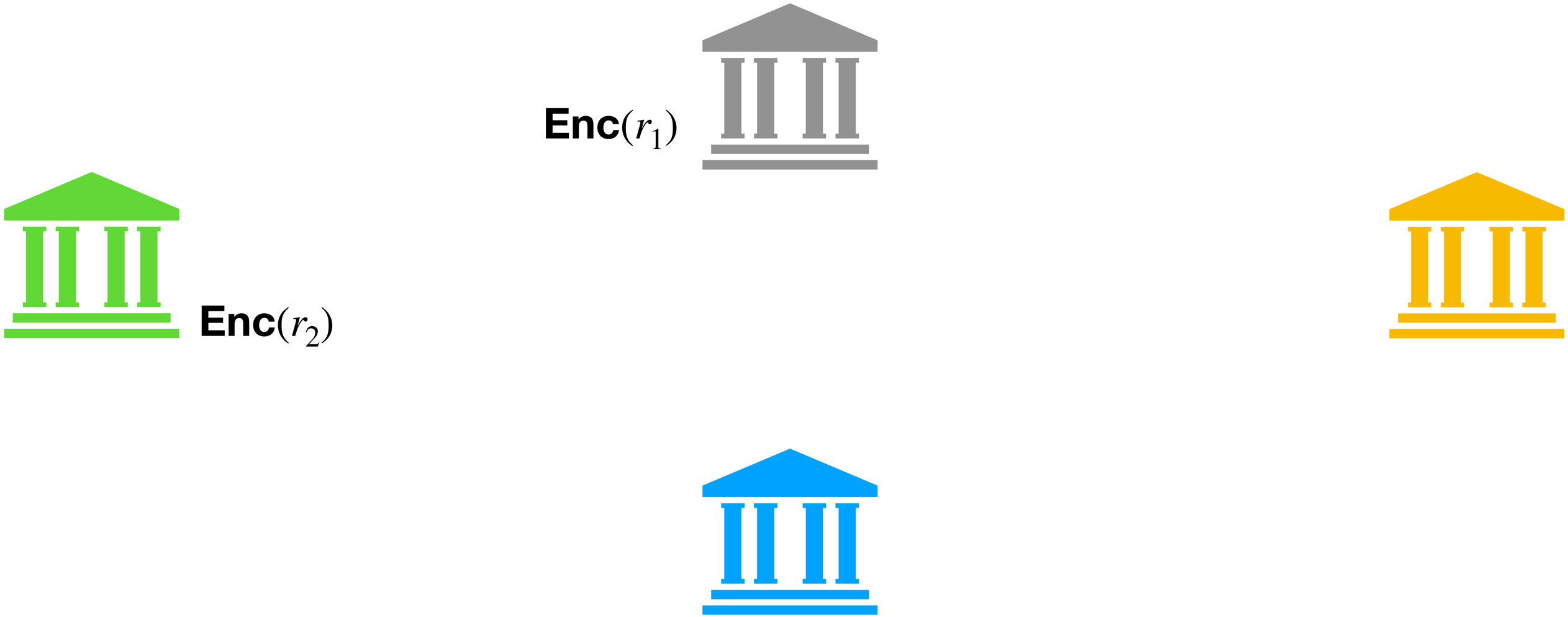
Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



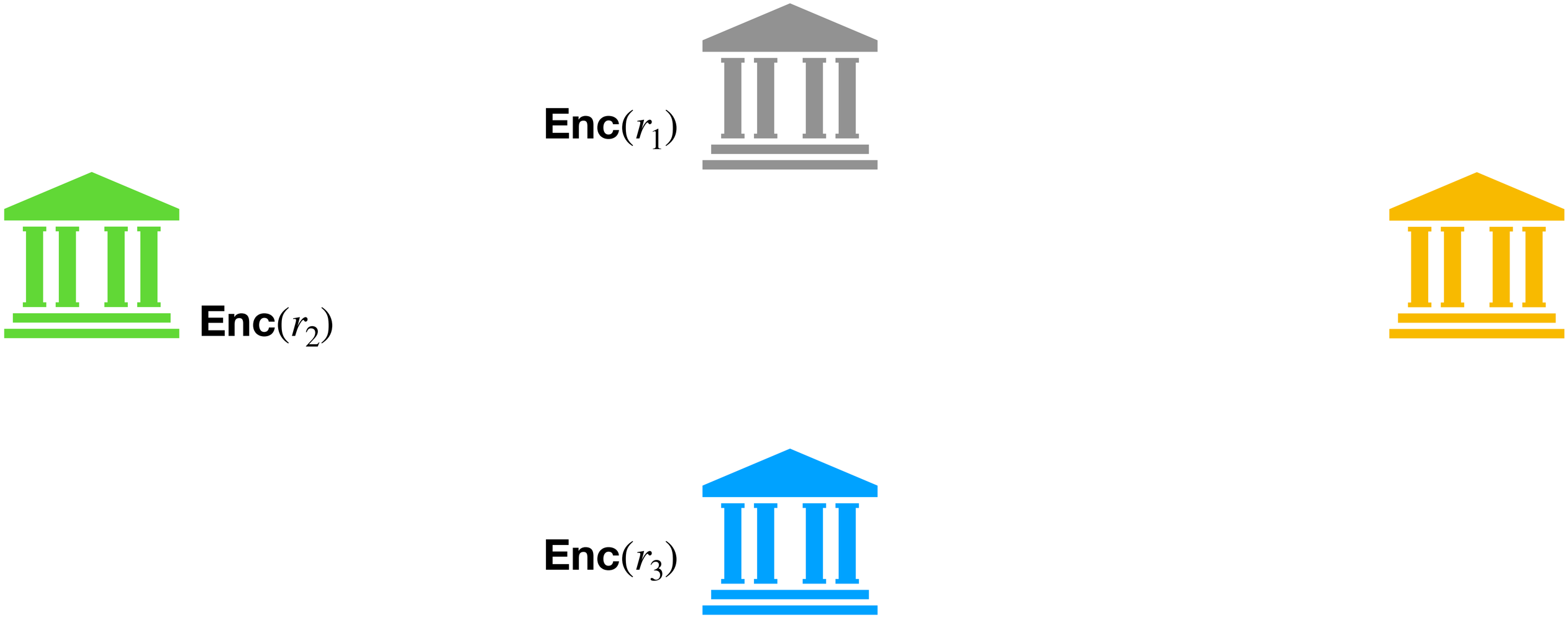
Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



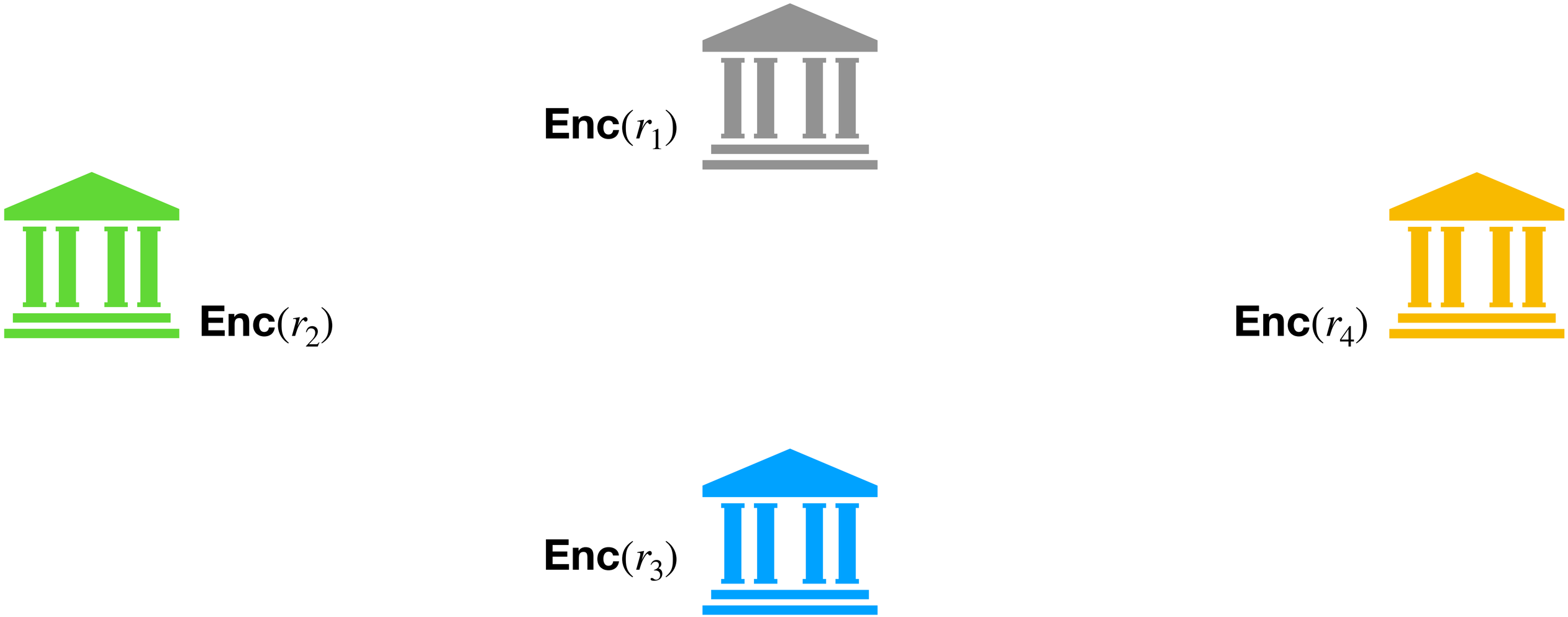
Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



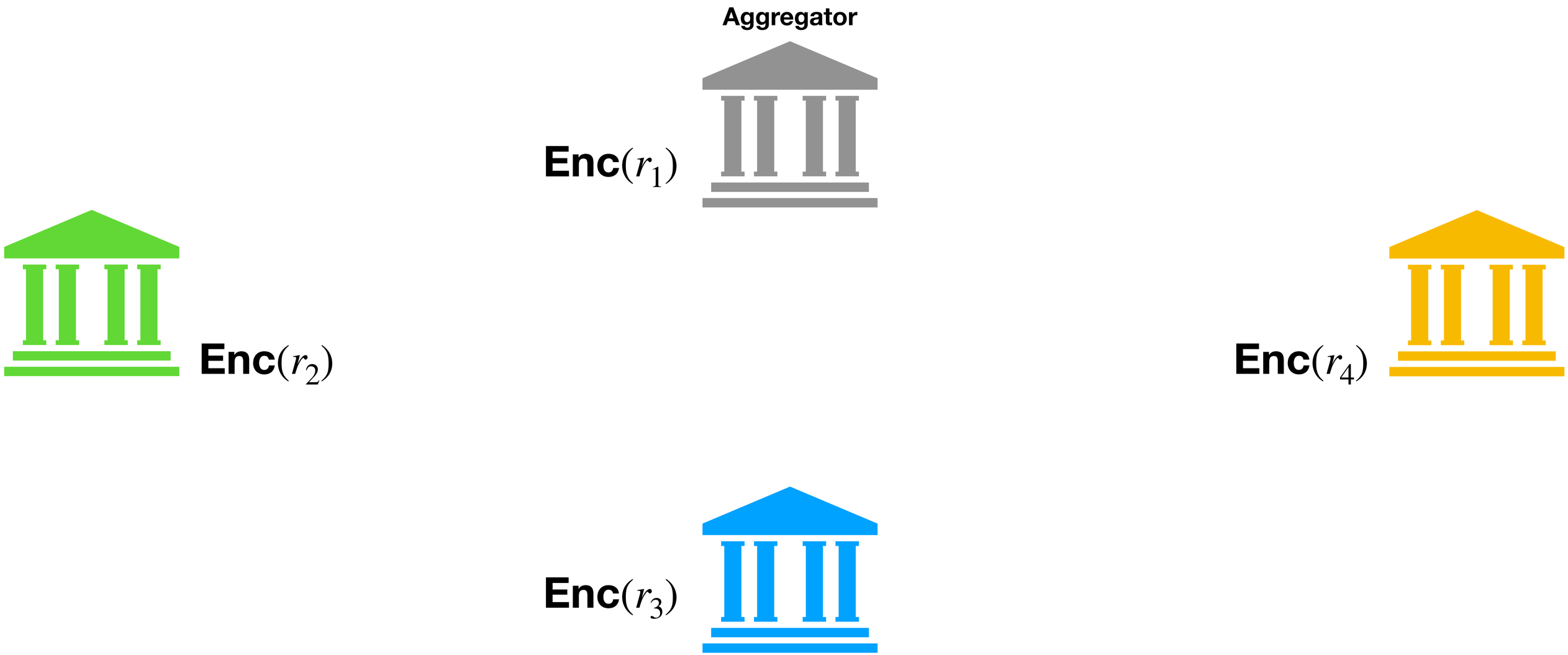
Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



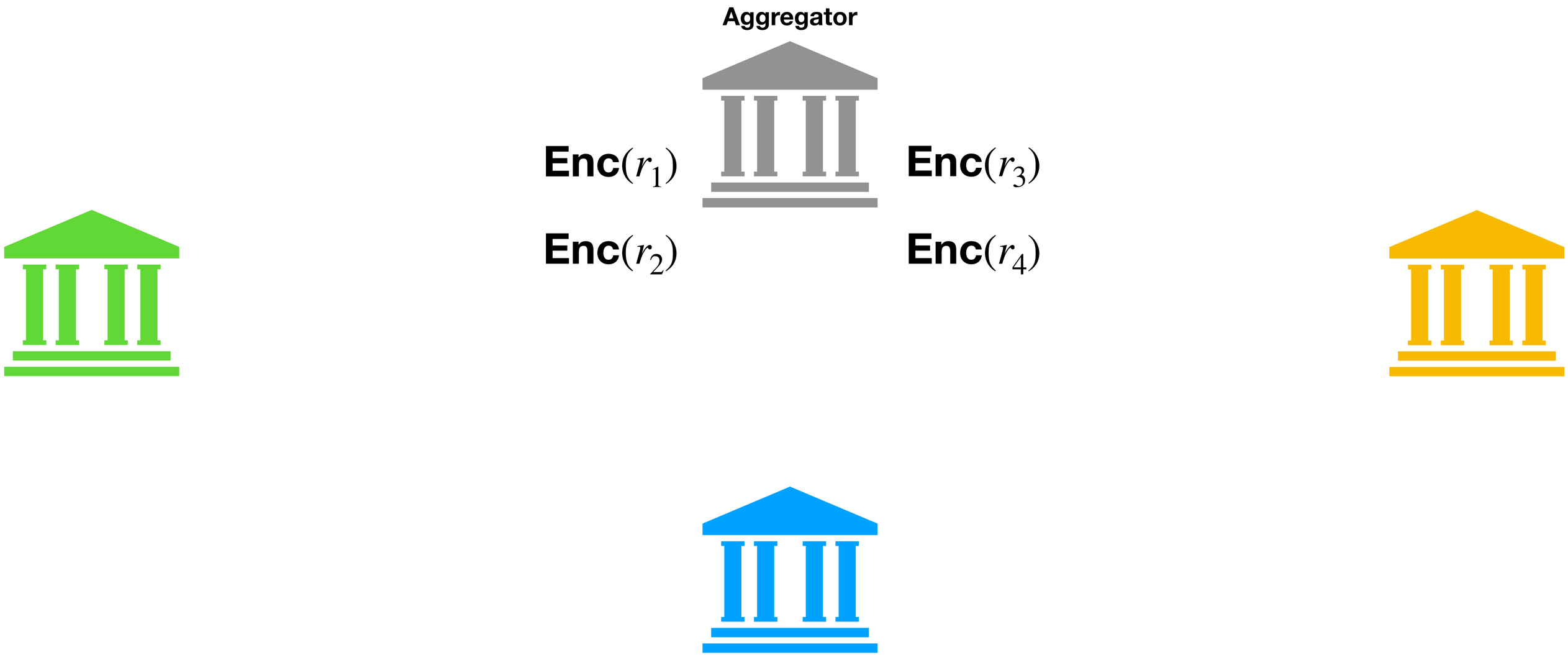
Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



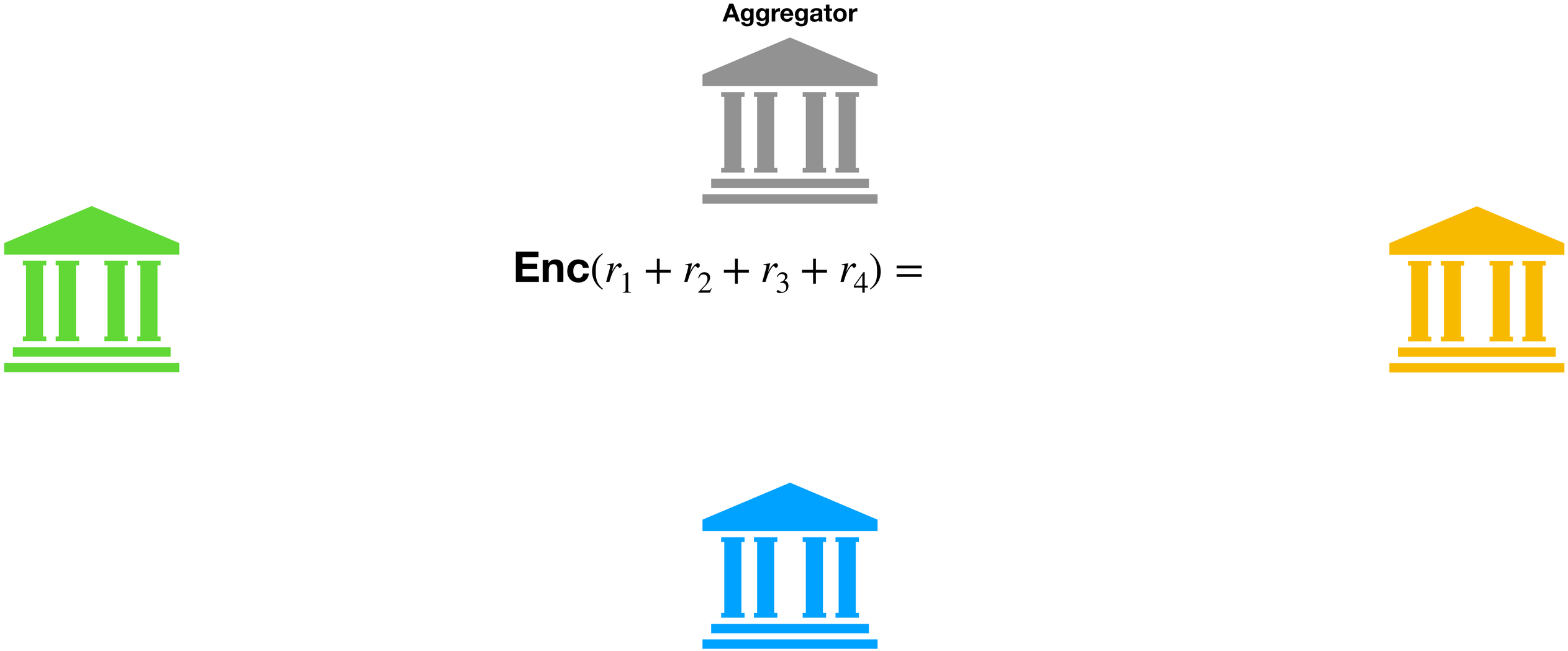
Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



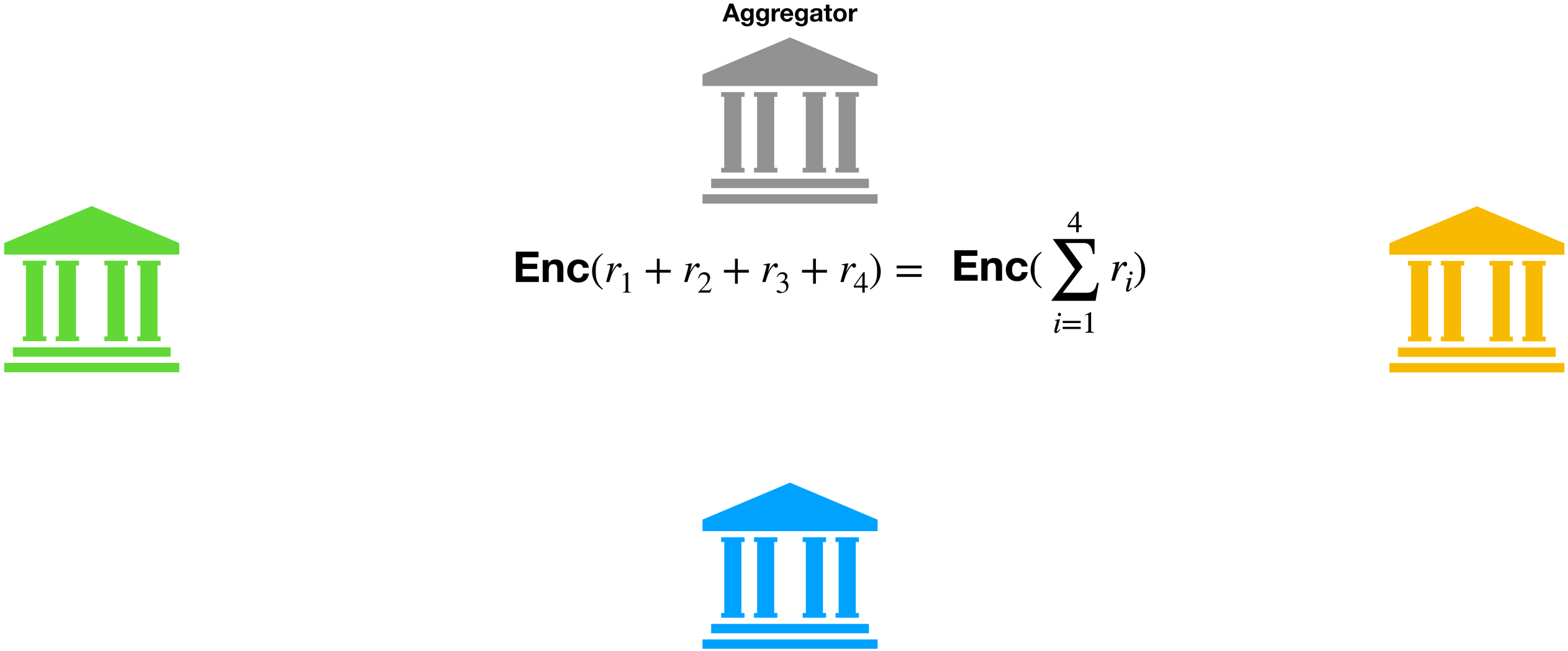
Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



Optimizing based on physical layout

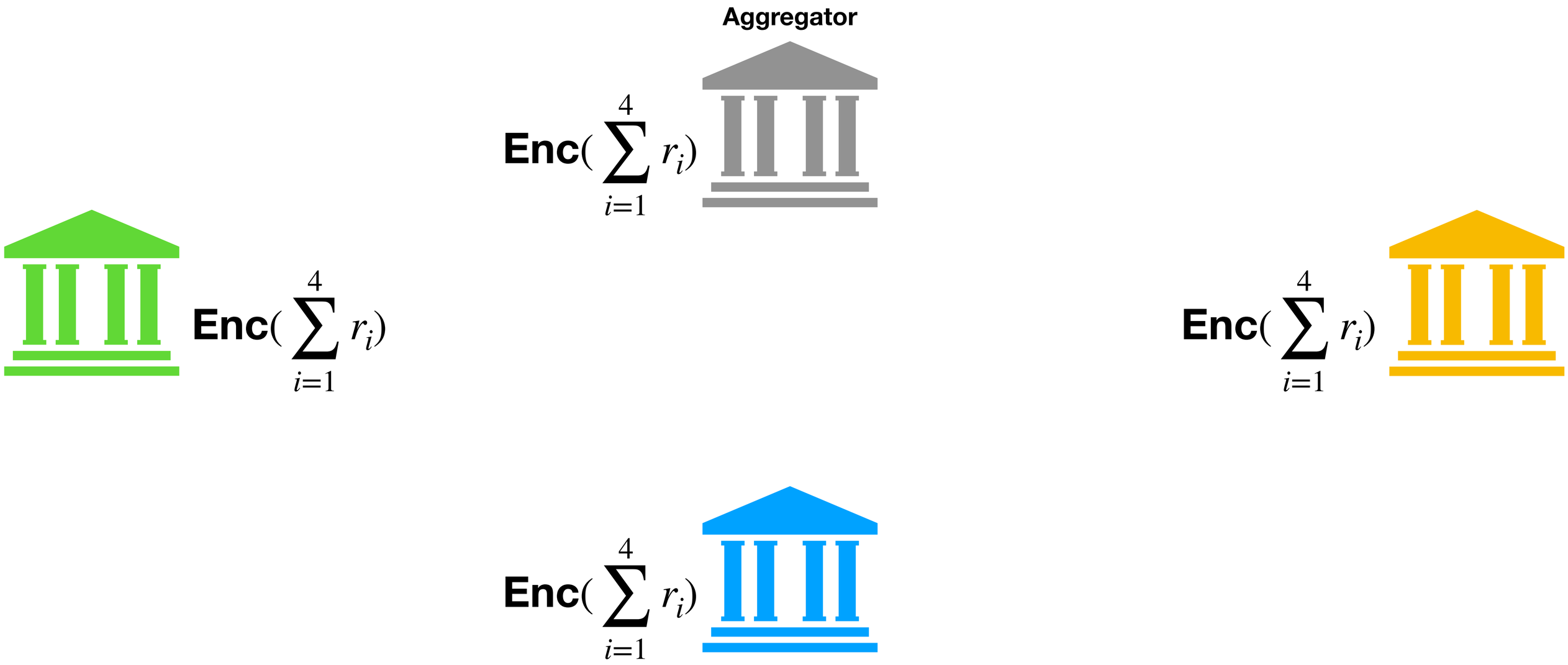
Example: Semi-Honest Triple Generation [DPSZ11]



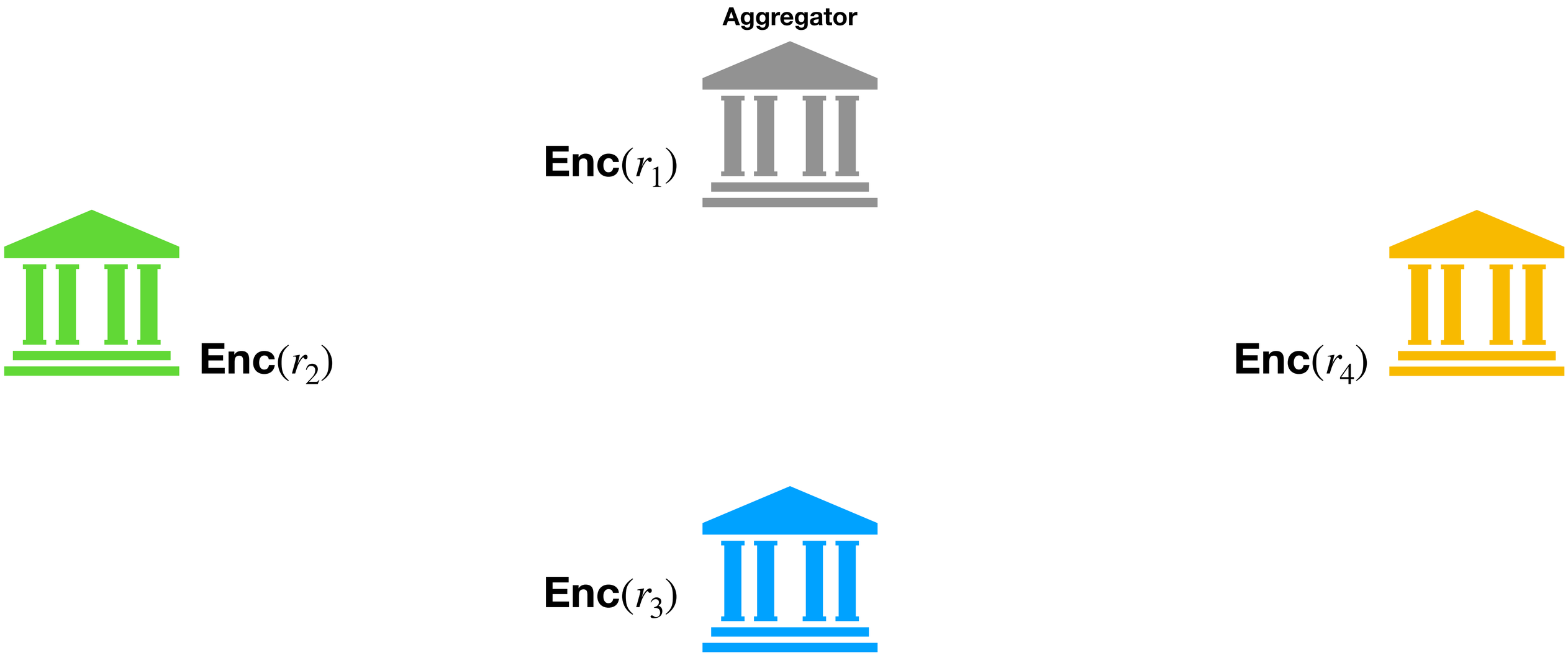
$$\textbf{Enc}(r_1 + r_2 + r_3 + r_4) = \textbf{Enc}(\sum_{i=1}^{4} r_i)$$

Optimizing based on physical layout

Example: Semi-Honest Triple Generation [DPSZ11]



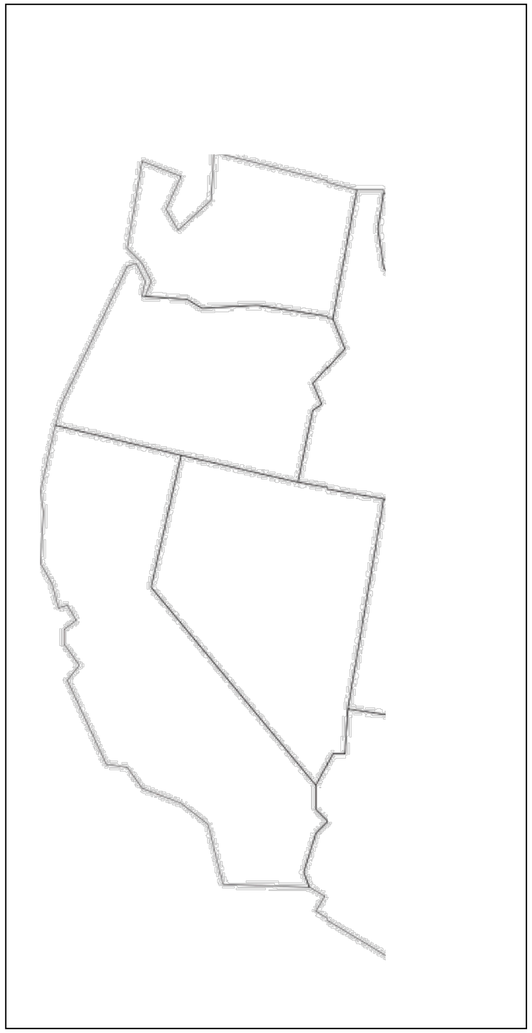
Optimizing based on physical layout

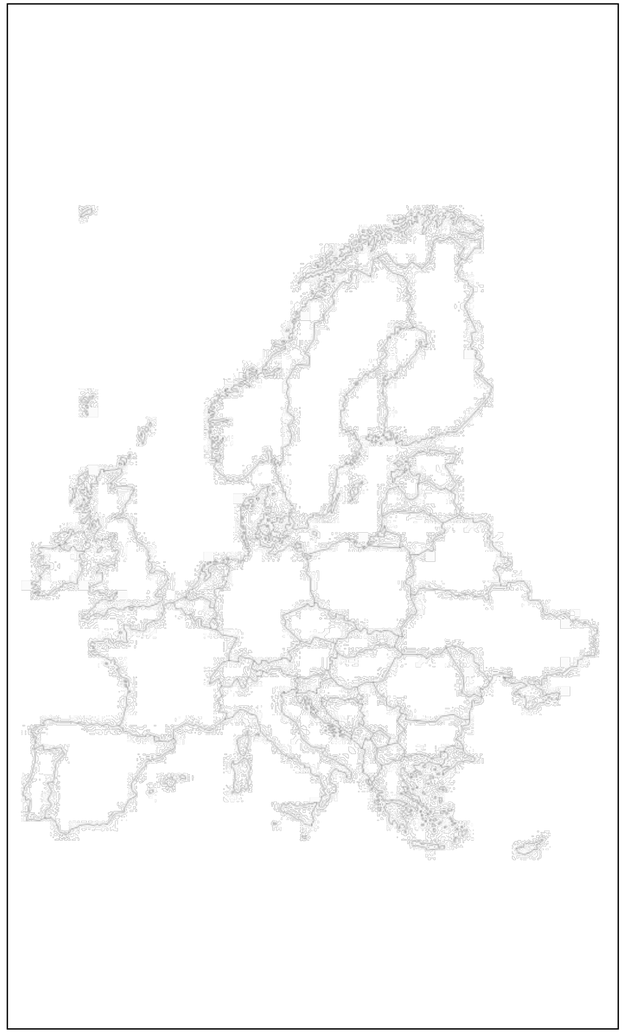


Optimizing based on physical layout

Seattle

$\mathbf{Enc}(r_2)$





$\mathbf{Enc}(r_1)$

Aggregator

London

Los Angeles

$\mathbf{Enc}(r_3)$

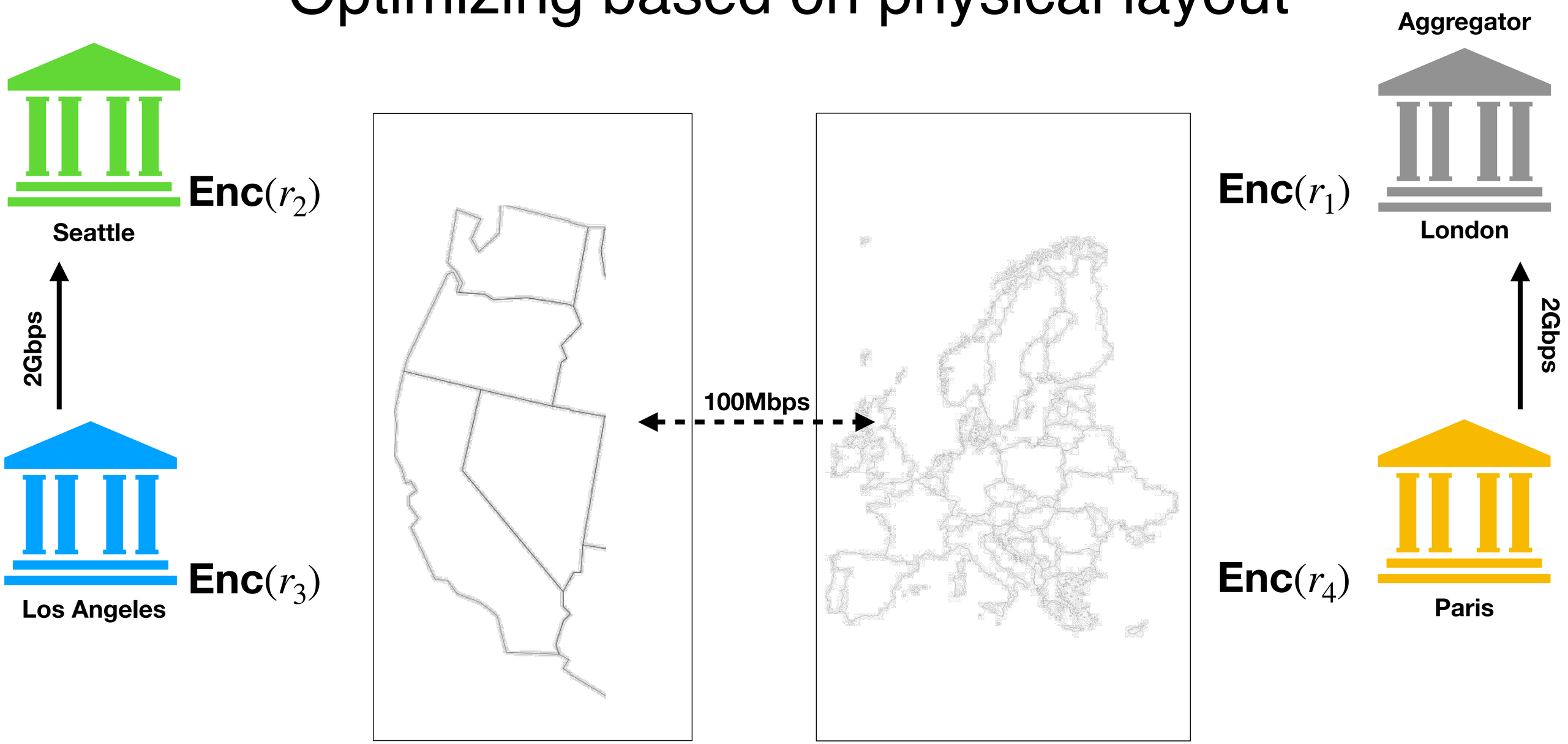
$\mathbf{Enc}(r_4)$

Paris

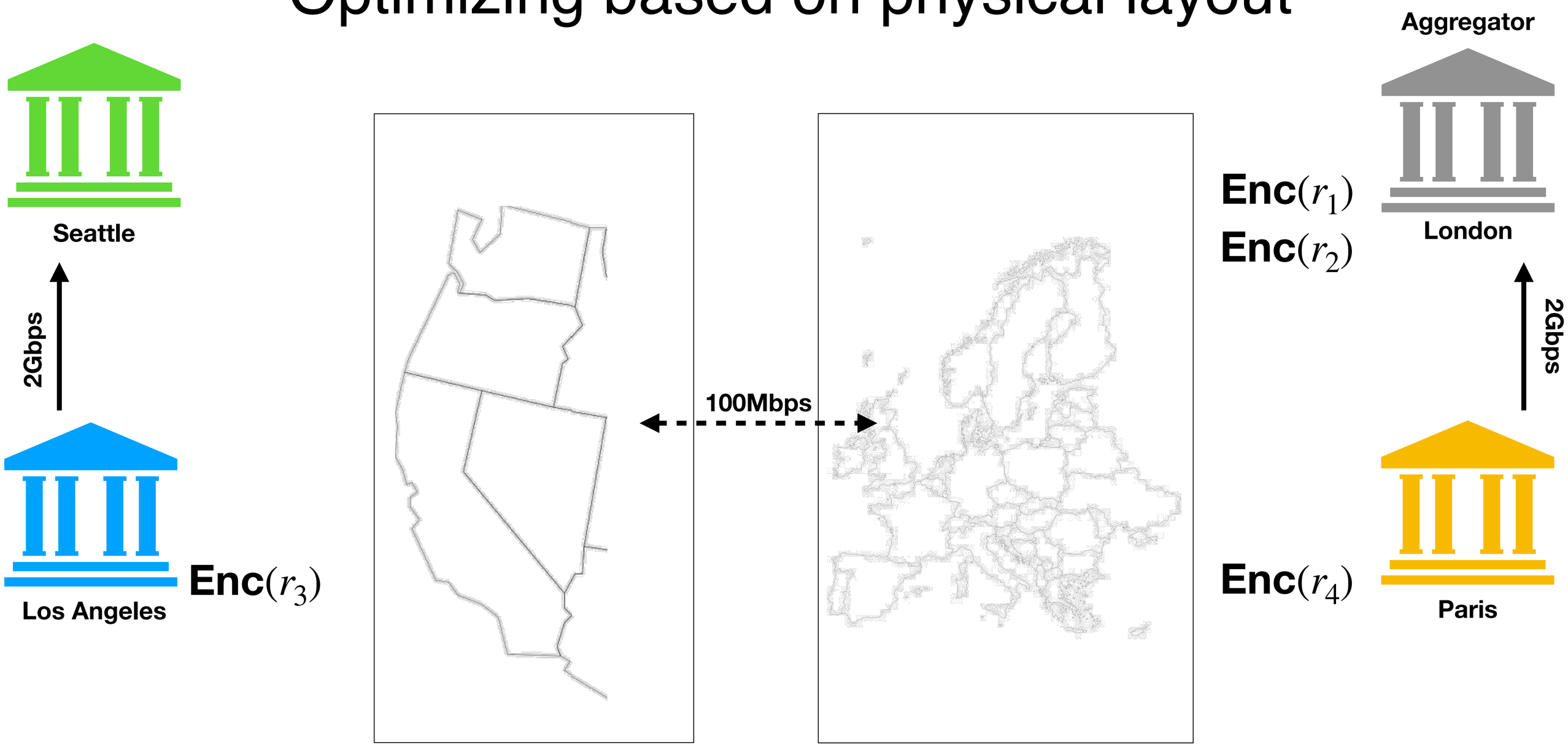
Optimizing based on physical layout

Optimizing based on physical layout

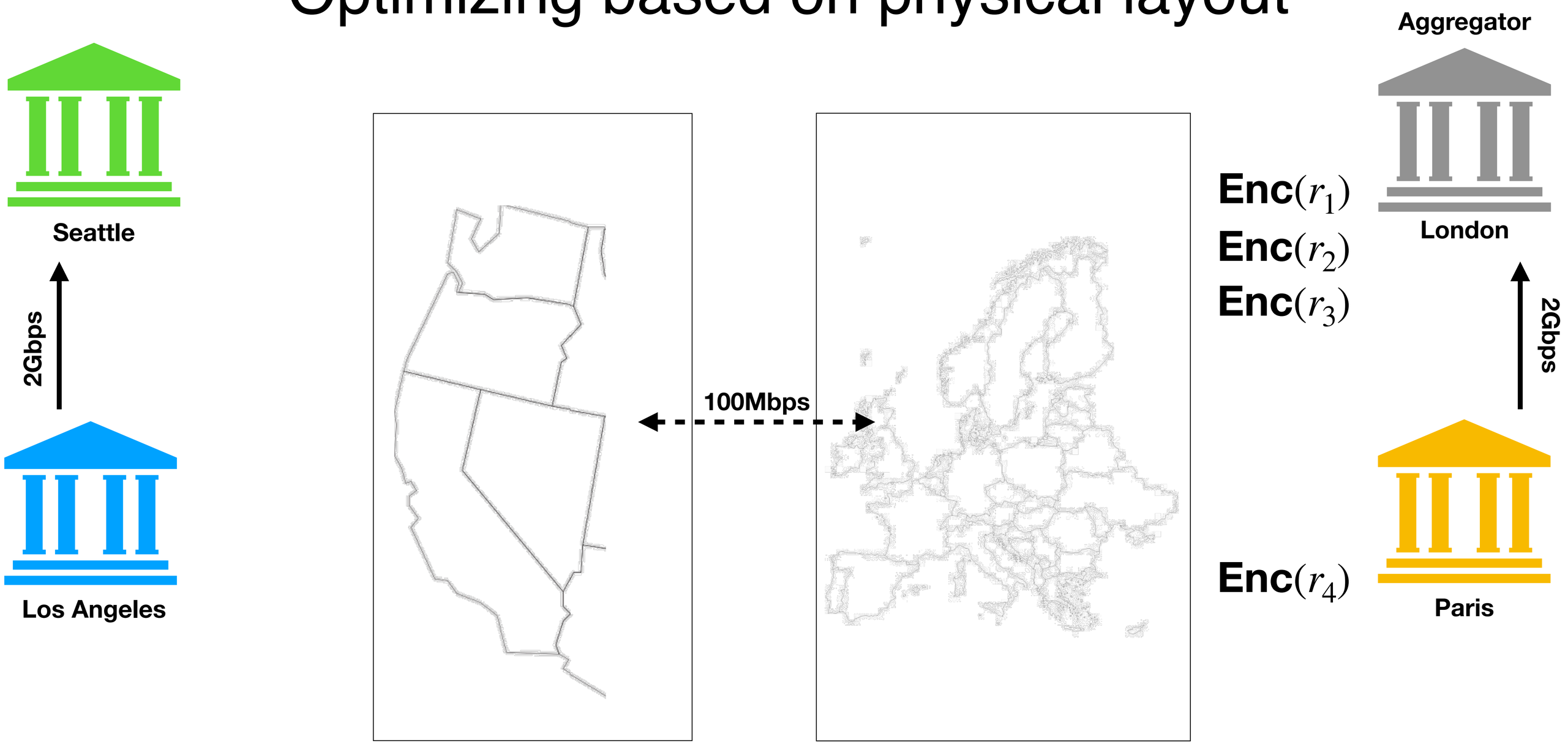
Optimizing based on physical layout



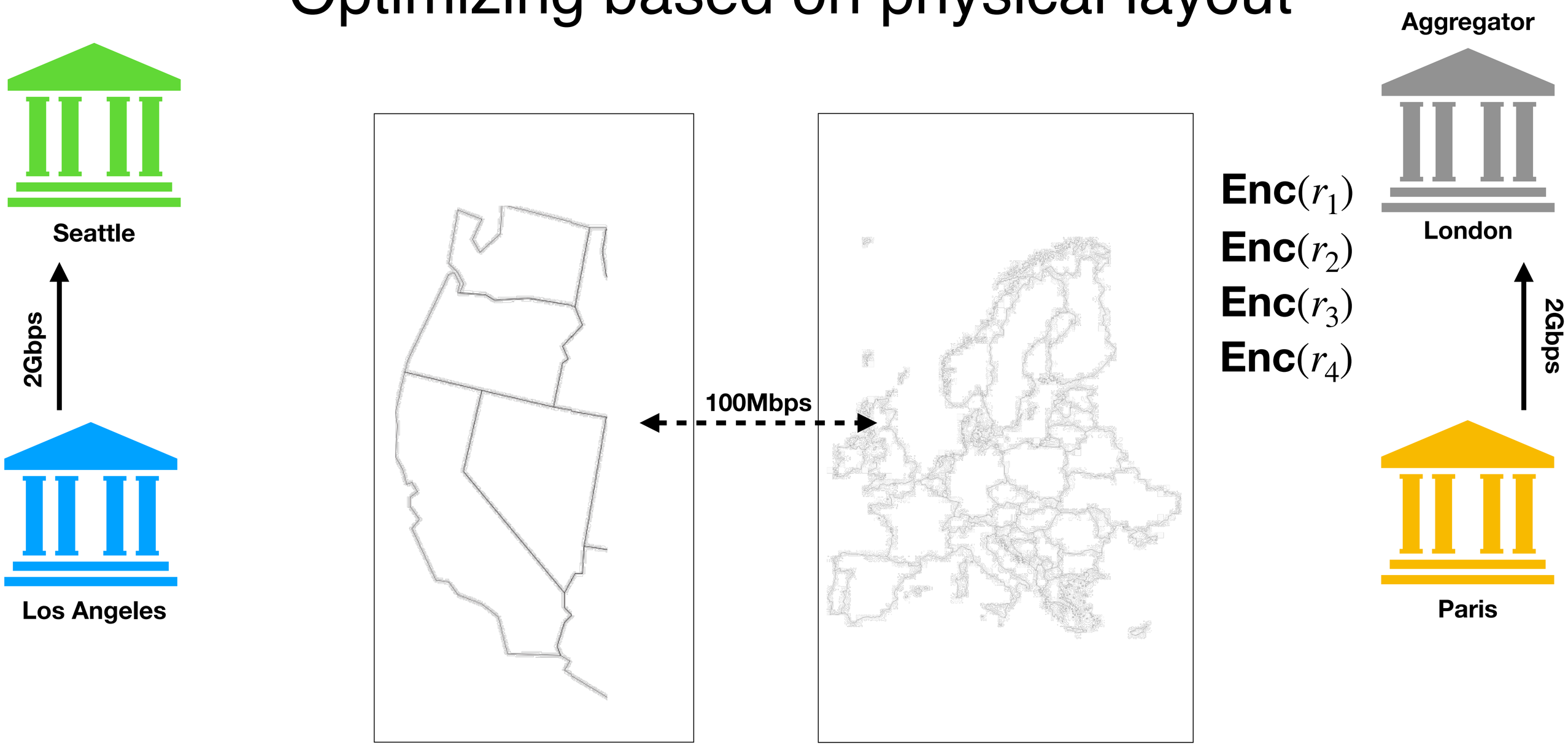
Optimizing based on physical layout



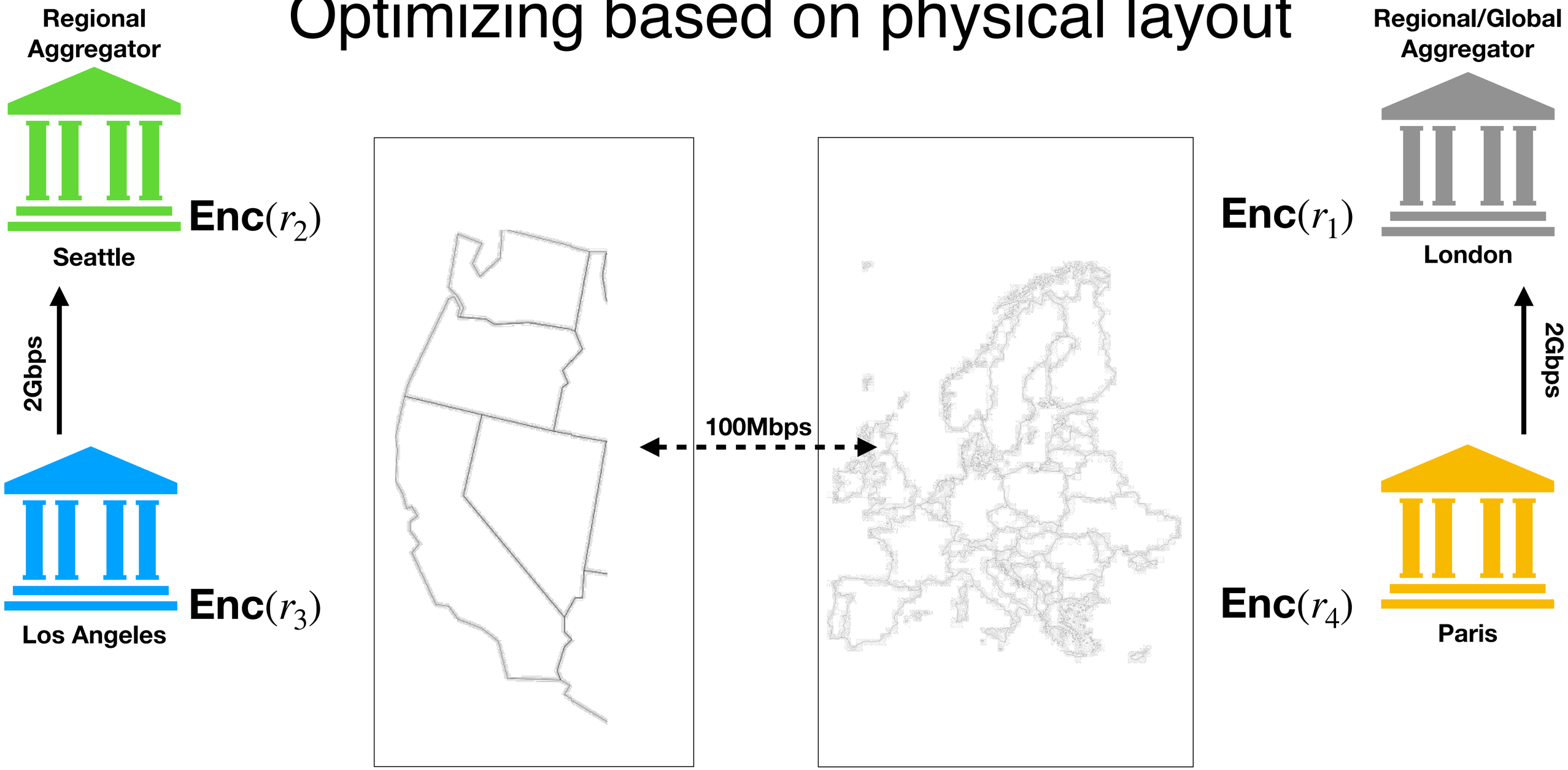
Optimizing based on physical layout



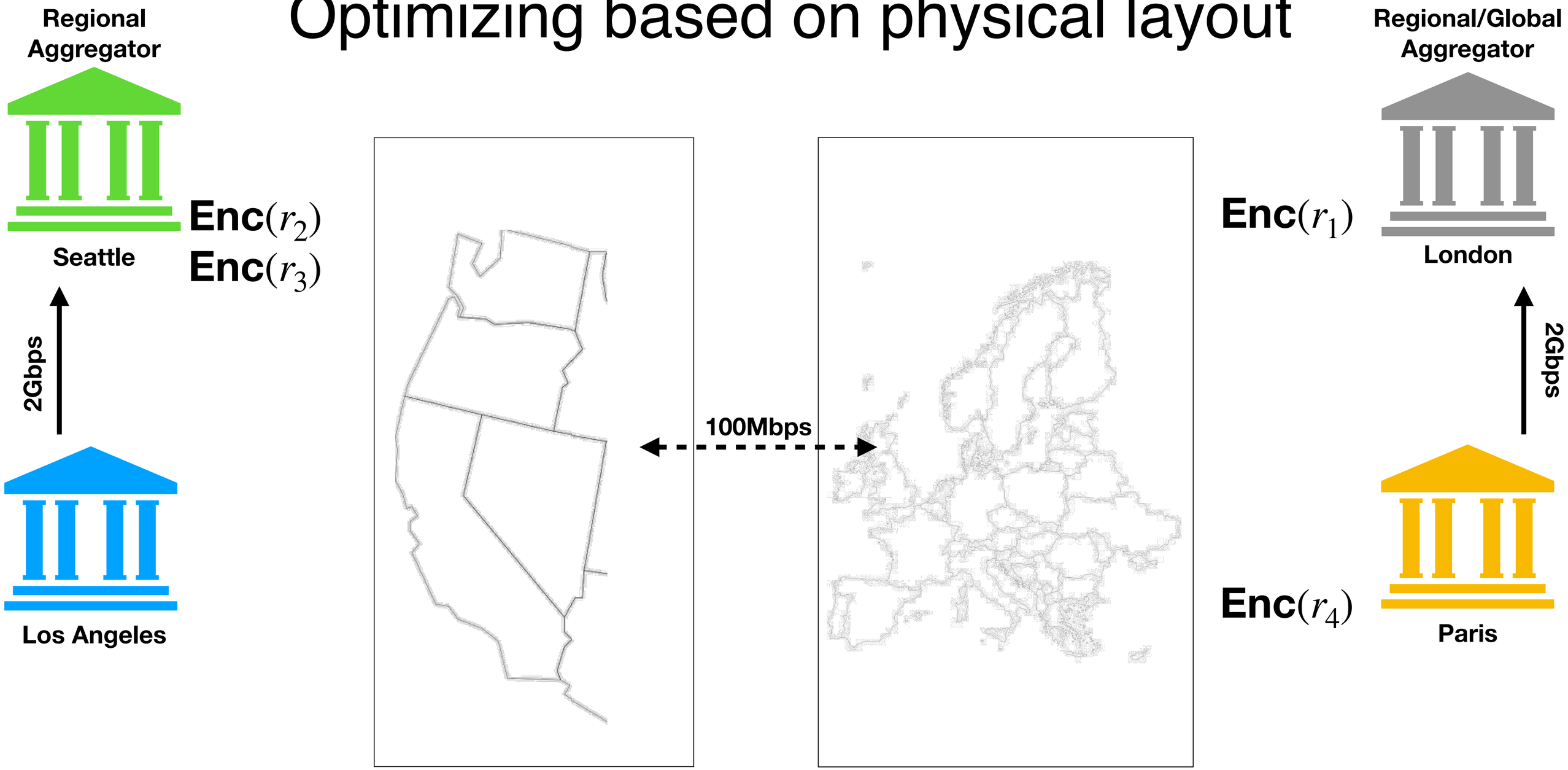
Optimizing based on physical layout



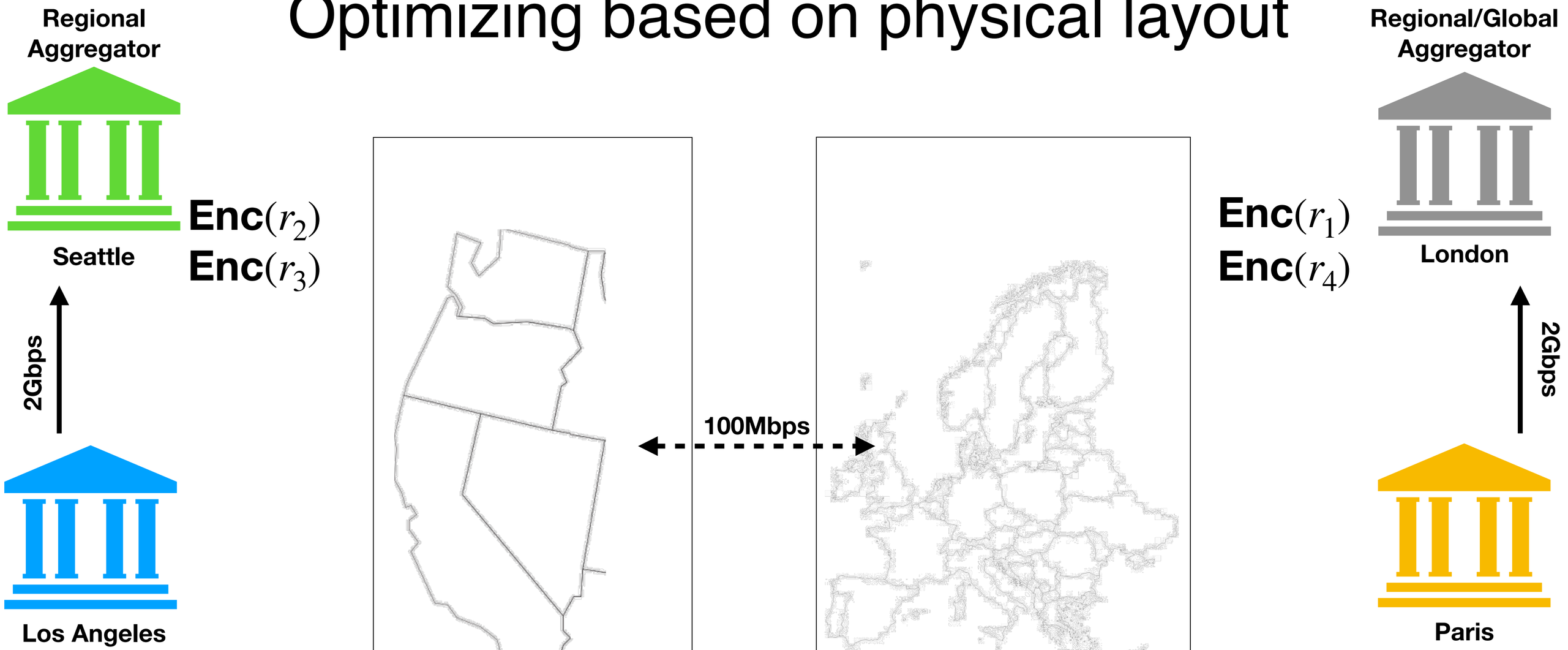
Optimizing based on physical layout



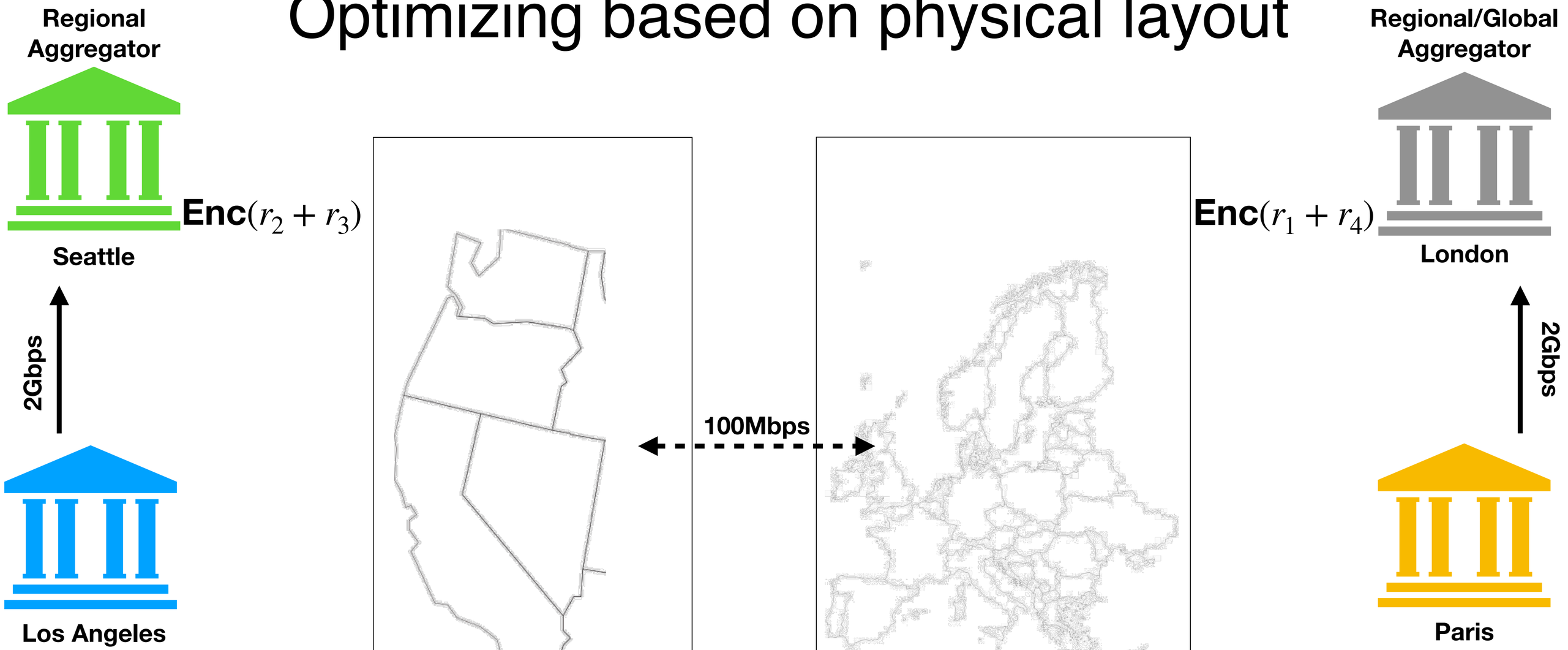
Optimizing based on physical layout



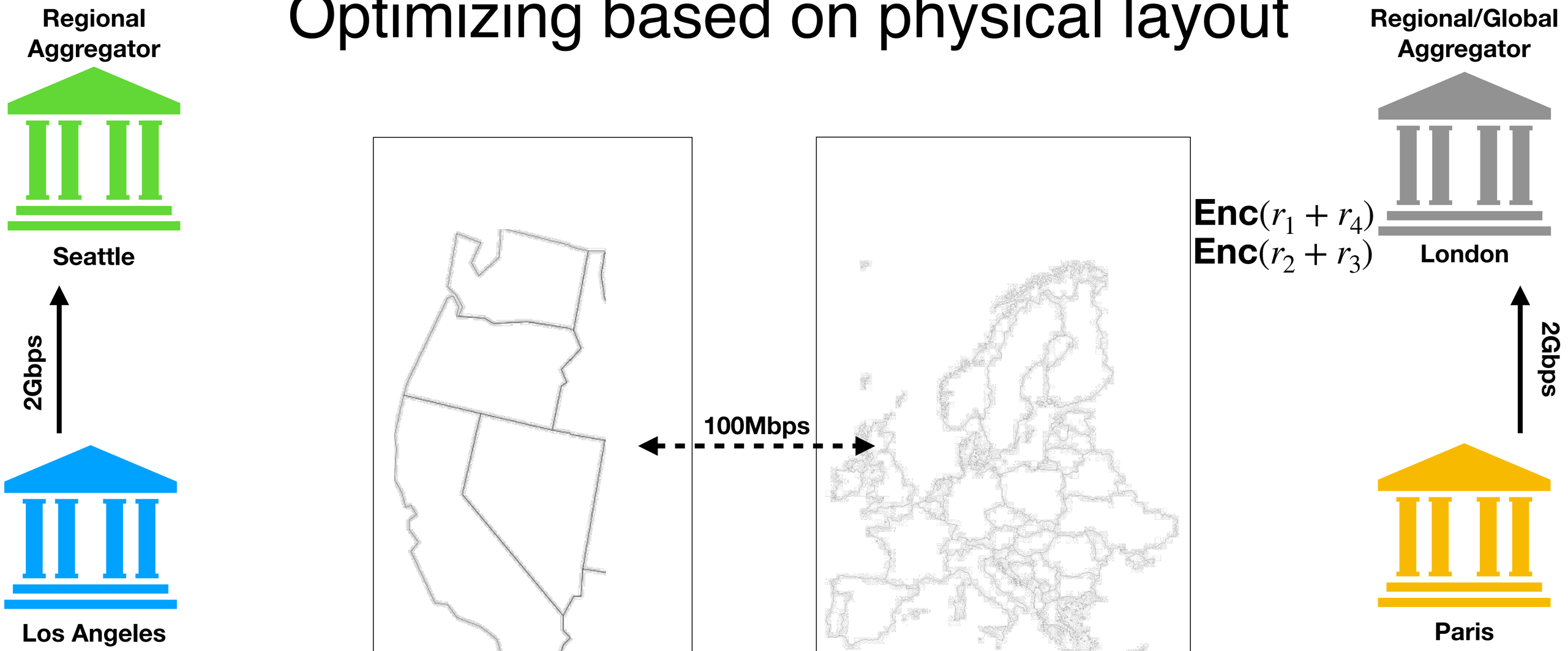
Optimizing based on physical layout



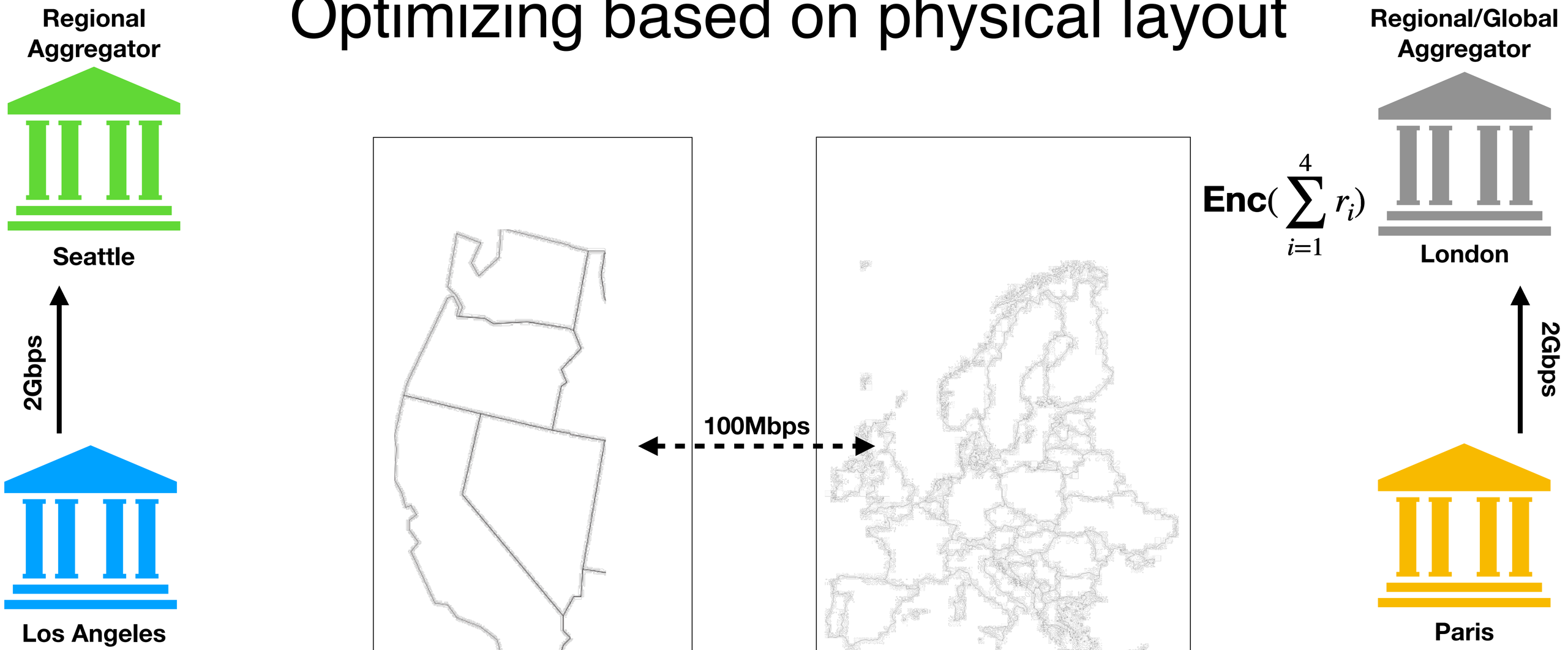
Optimizing based on physical layout



Optimizing based on physical layout

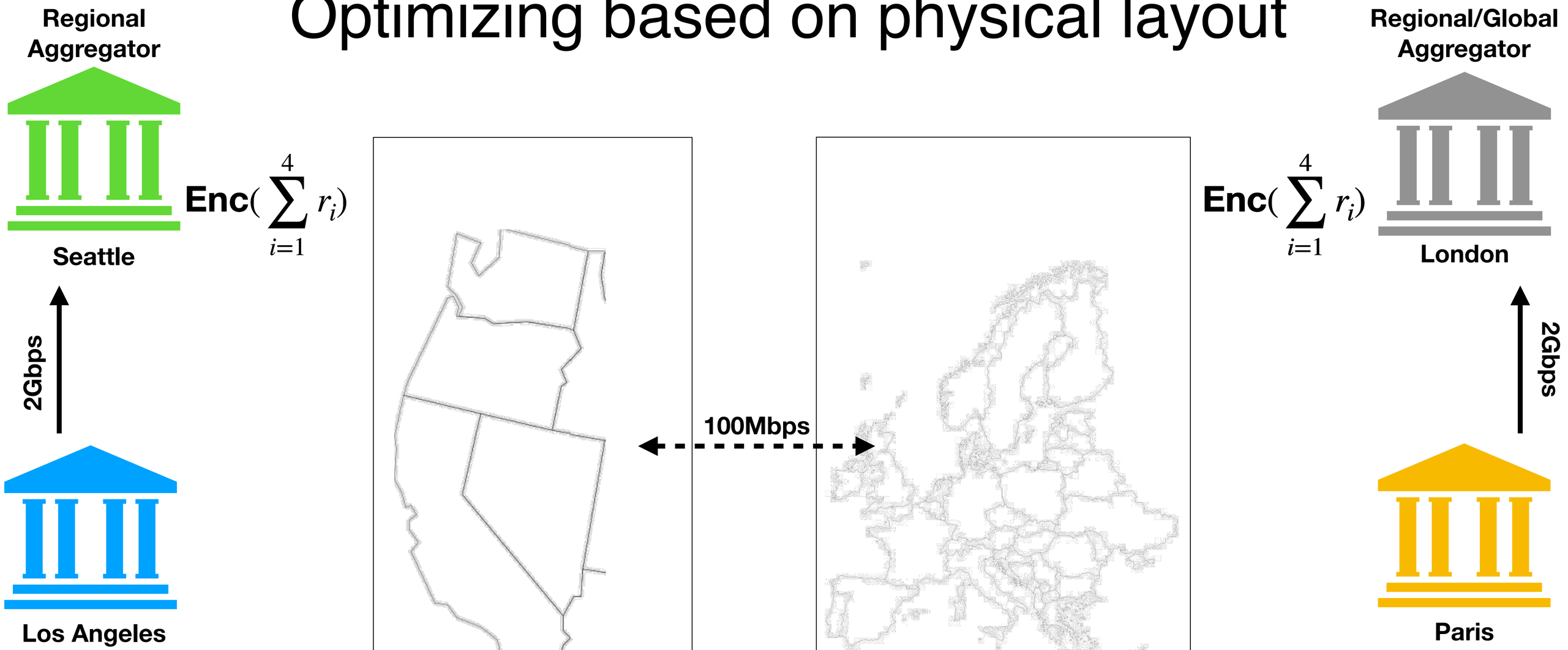


Optimizing based on physical layout

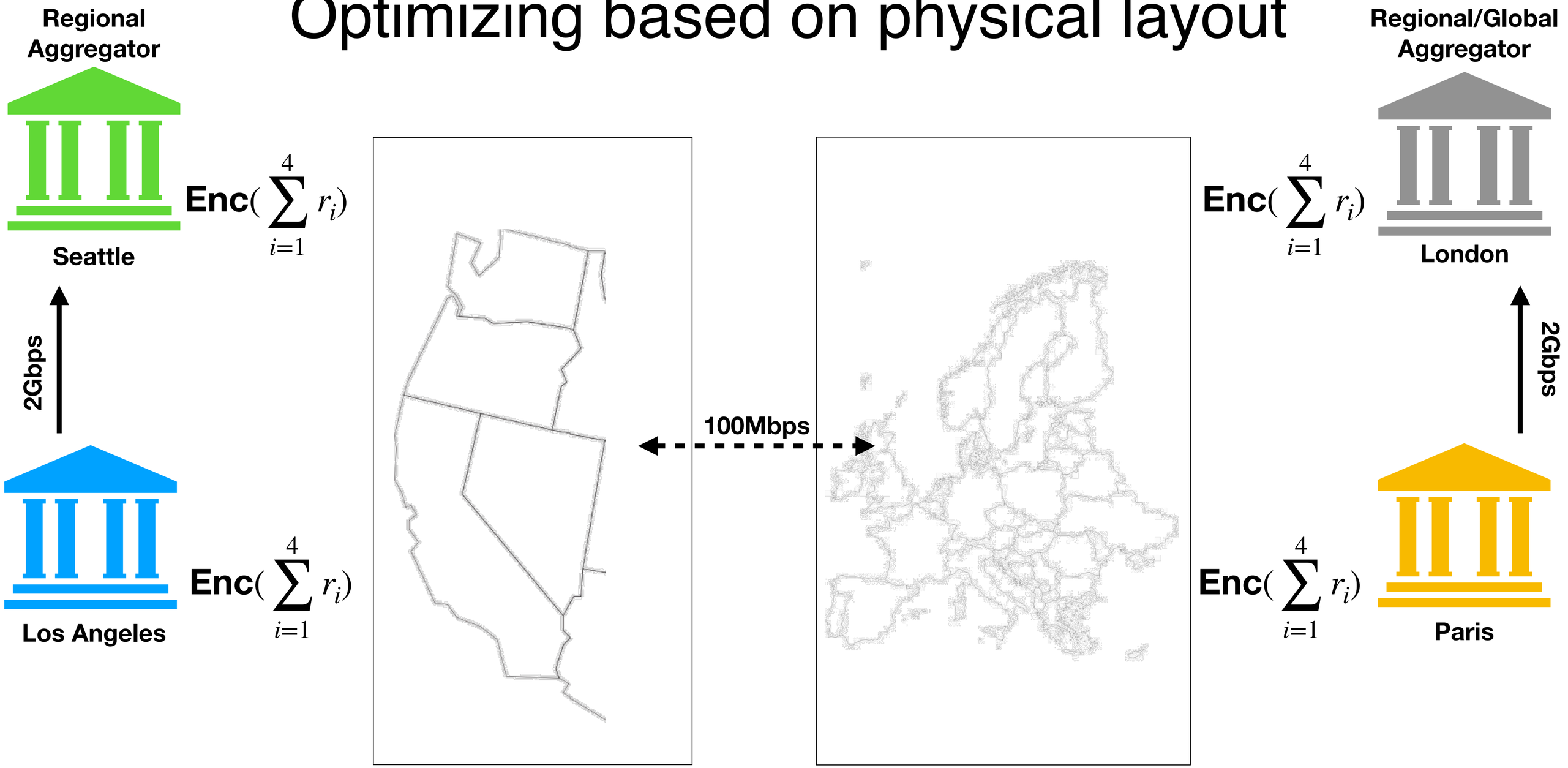


$$\mathrm{Enc}\left(\sum_{i=1}^{4} r_i\right)$$

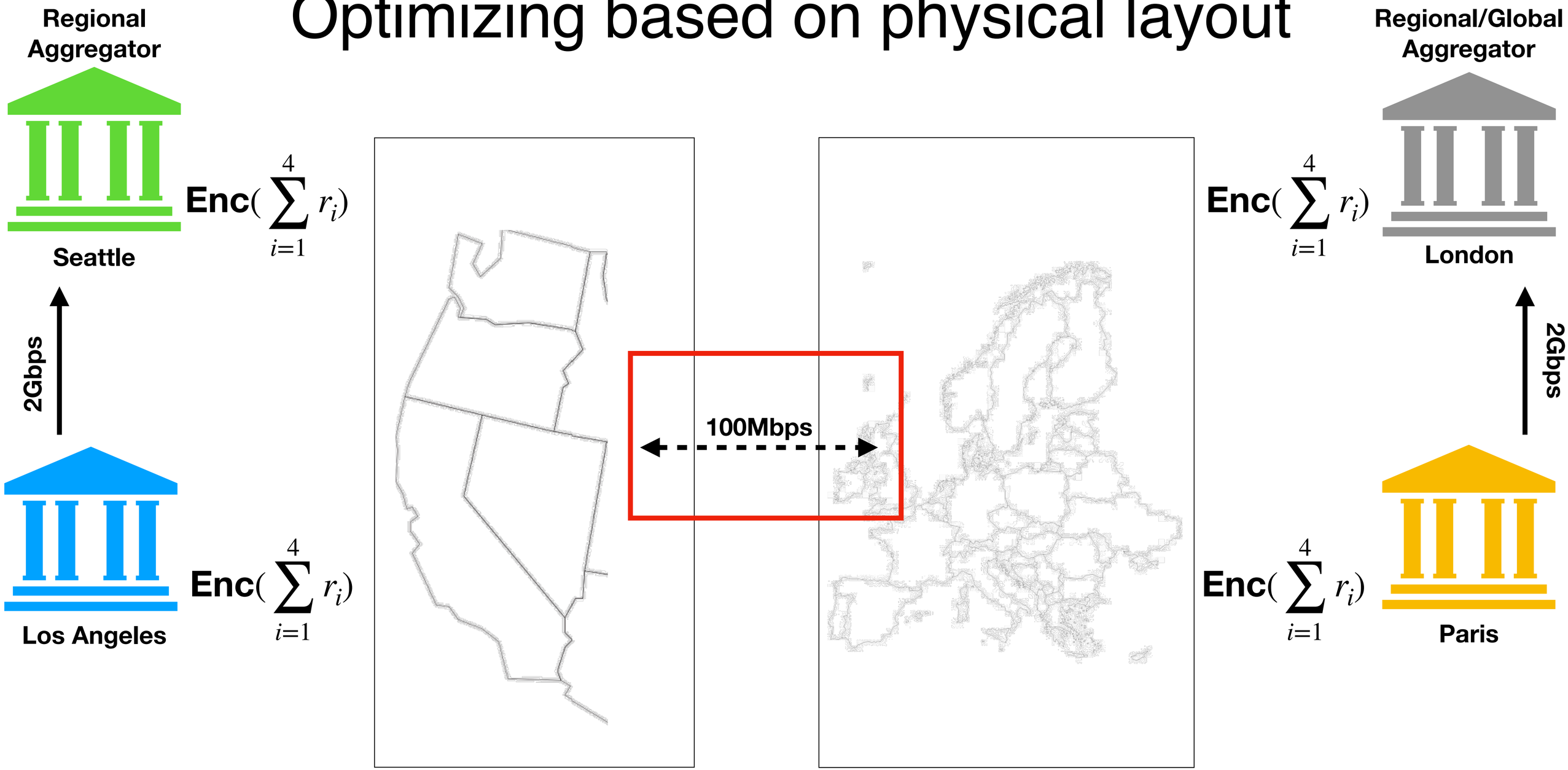
Optimizing based on physical layout



Optimizing based on physical layout



Optimizing based on physical layout



Evaluation

Impact of Cerebro’s Physical Layout Optimizations

Impact of Cerebro's Physical Layout Optimizations

- 12 parties, 2 regions

Impact of Cerebro's Physical Layout Optimizations

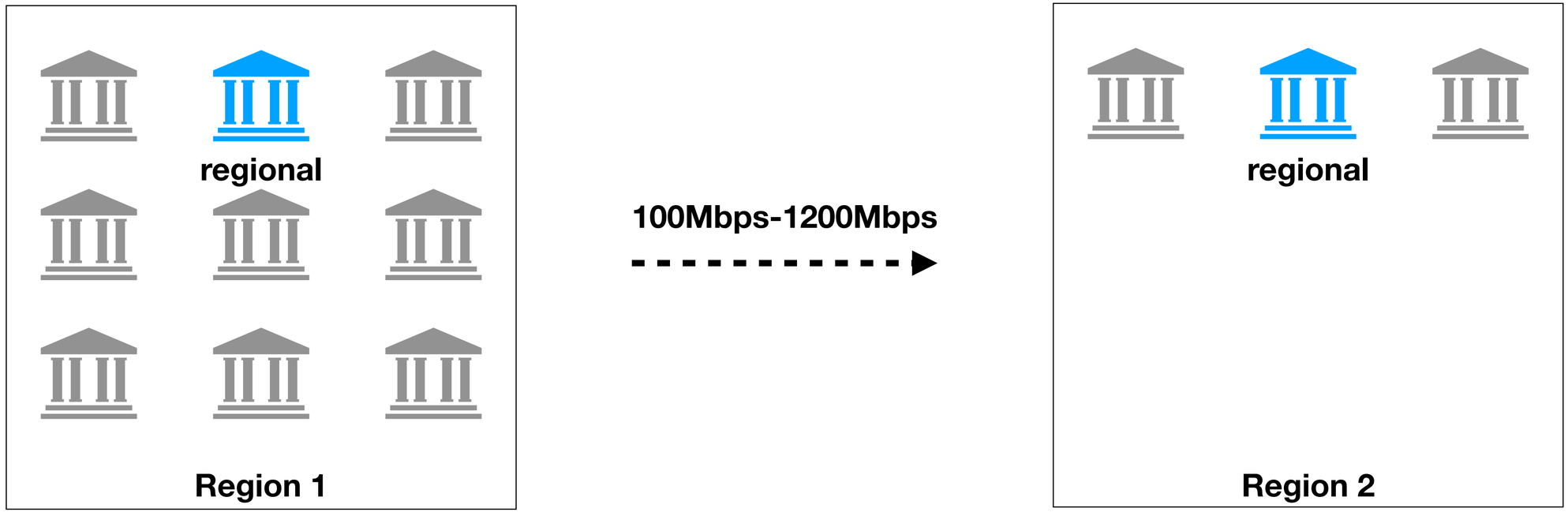
- 12 parties, 2 regions
  - 9-3 split

Impact of Cerebro's Physical Layout Optimizations

- 12 parties, 2 regions
  - 9-3 split
  - 2Gbps connection within each region, varied total bandwidth across regions.

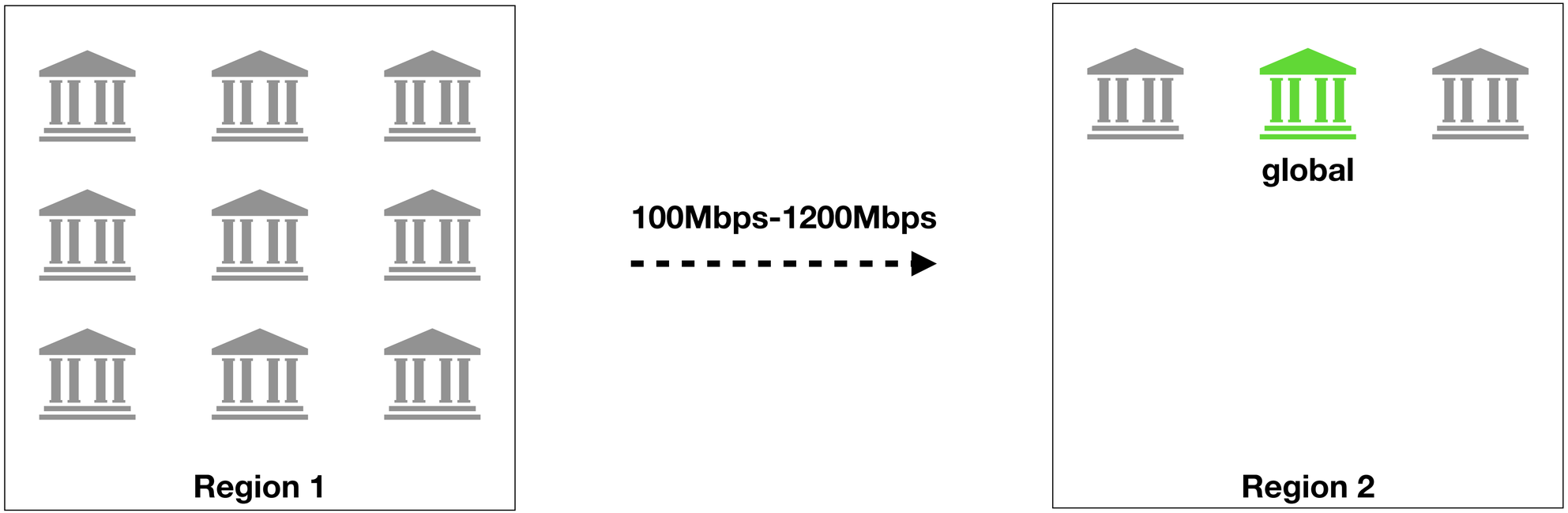
Impact of Cerebro's Physical Layout Optimizations

- 12 parties, 2 regions
  - 9-3 split
  - 2Gbps connection within each region, varied total bandwidth across regions.



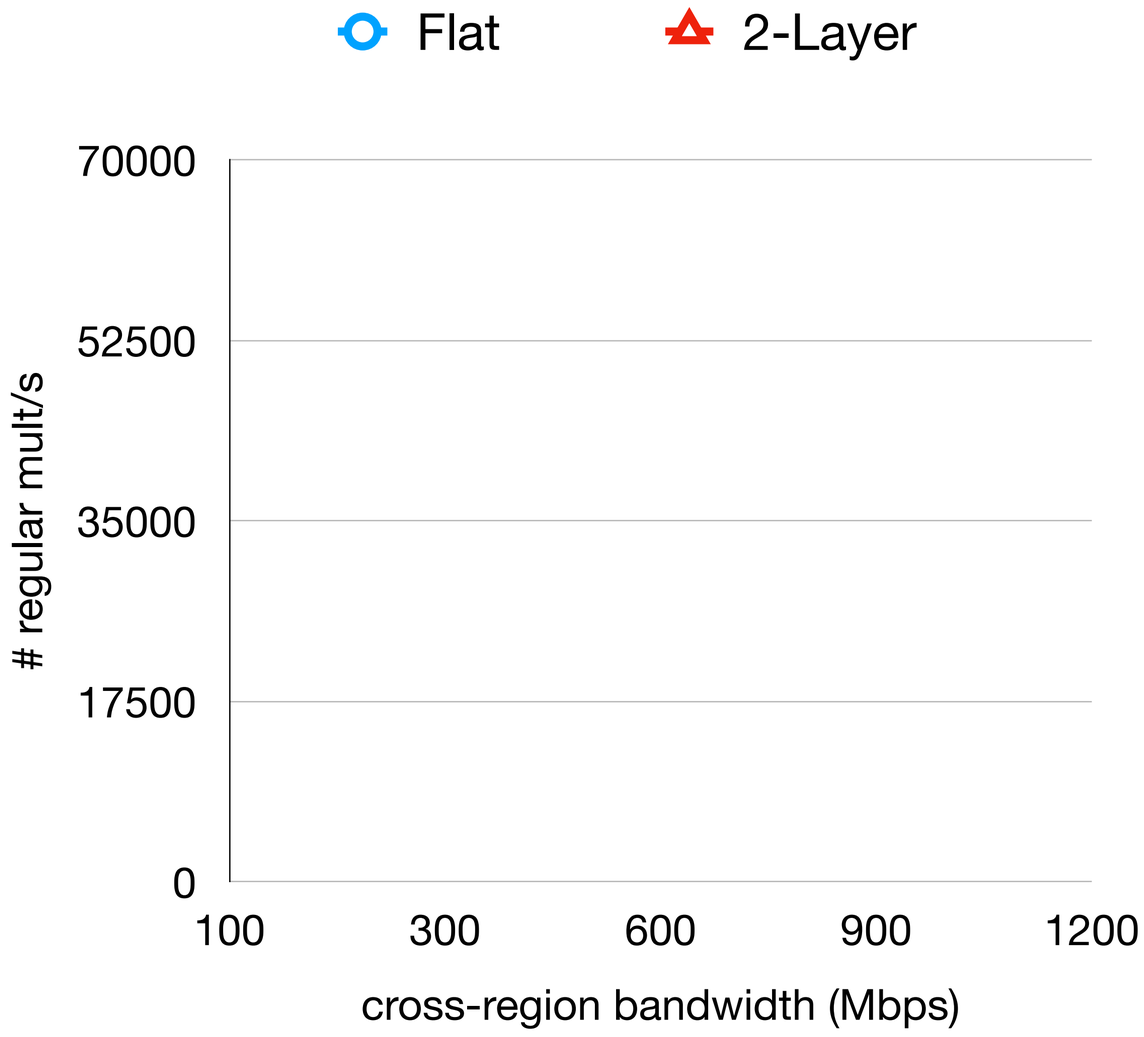
Impact of Cerebro's Physical Layout Optimizations

- 12 parties, 2 regions
  - 9-3 split
  - 2Gbps connection within each region, varied total bandwidth across regions.

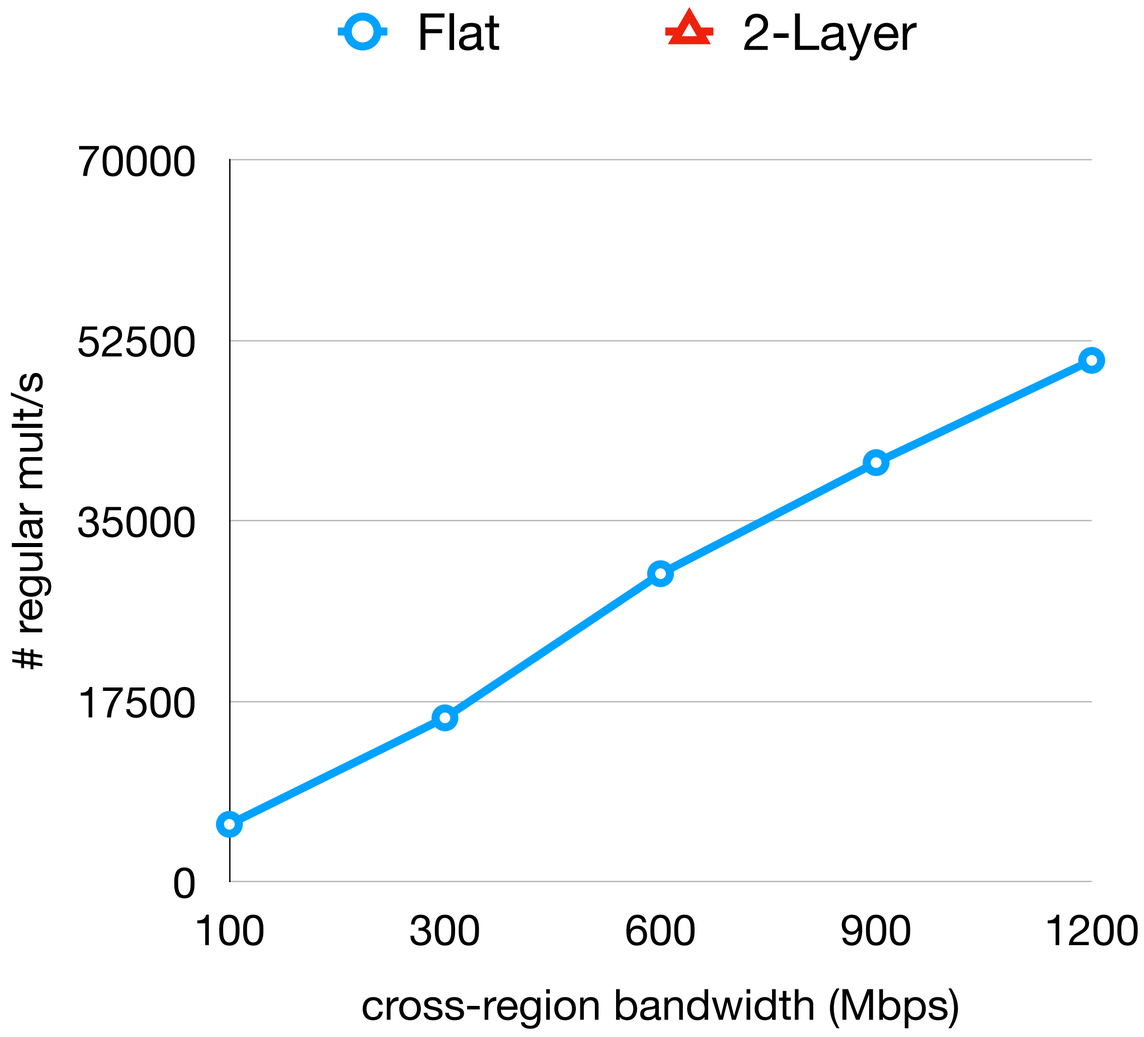


Impact of Cerebro’s Physical Layout Optimizations

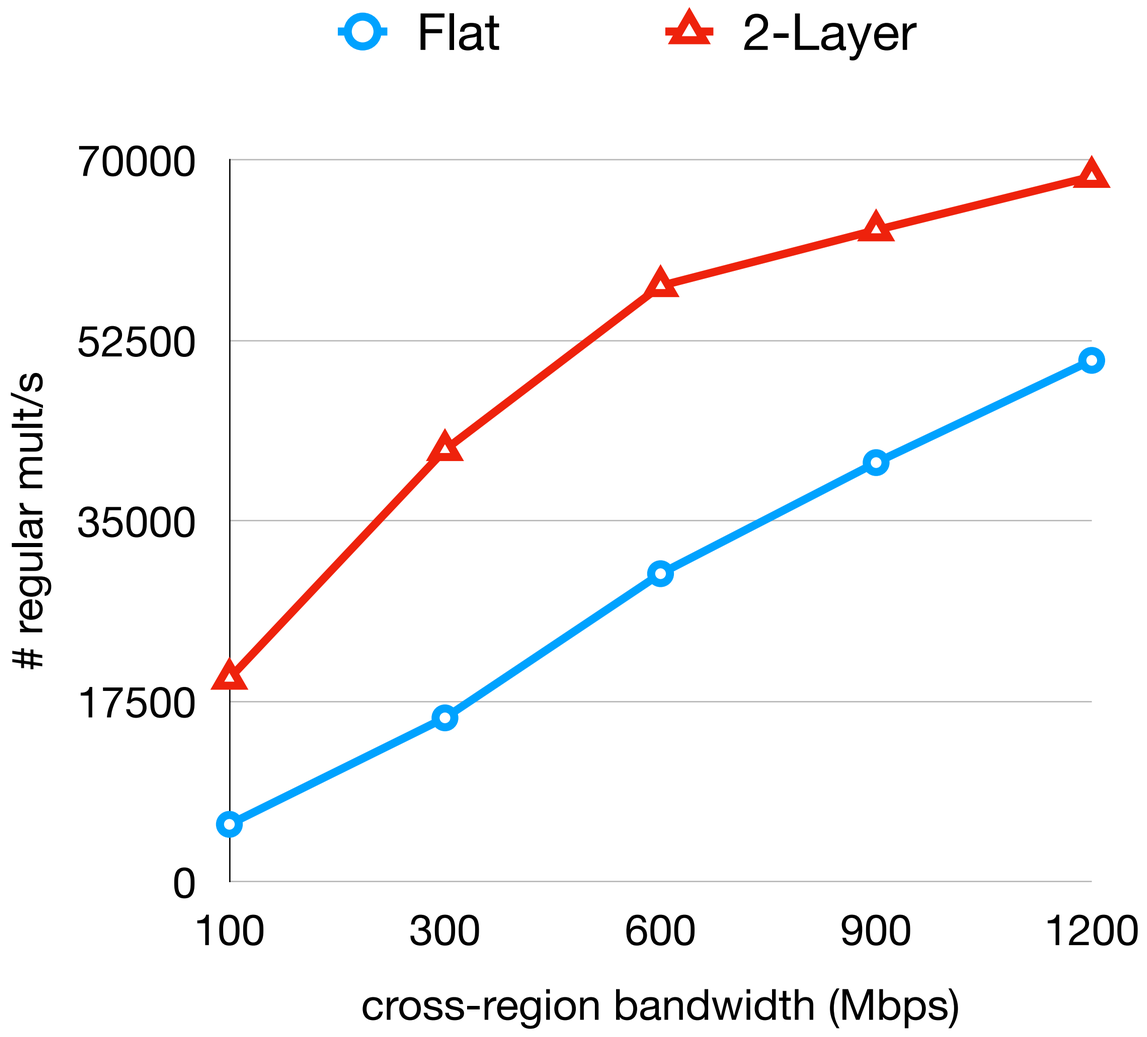
Impact of Cerebro's Physical Layout Optimizations



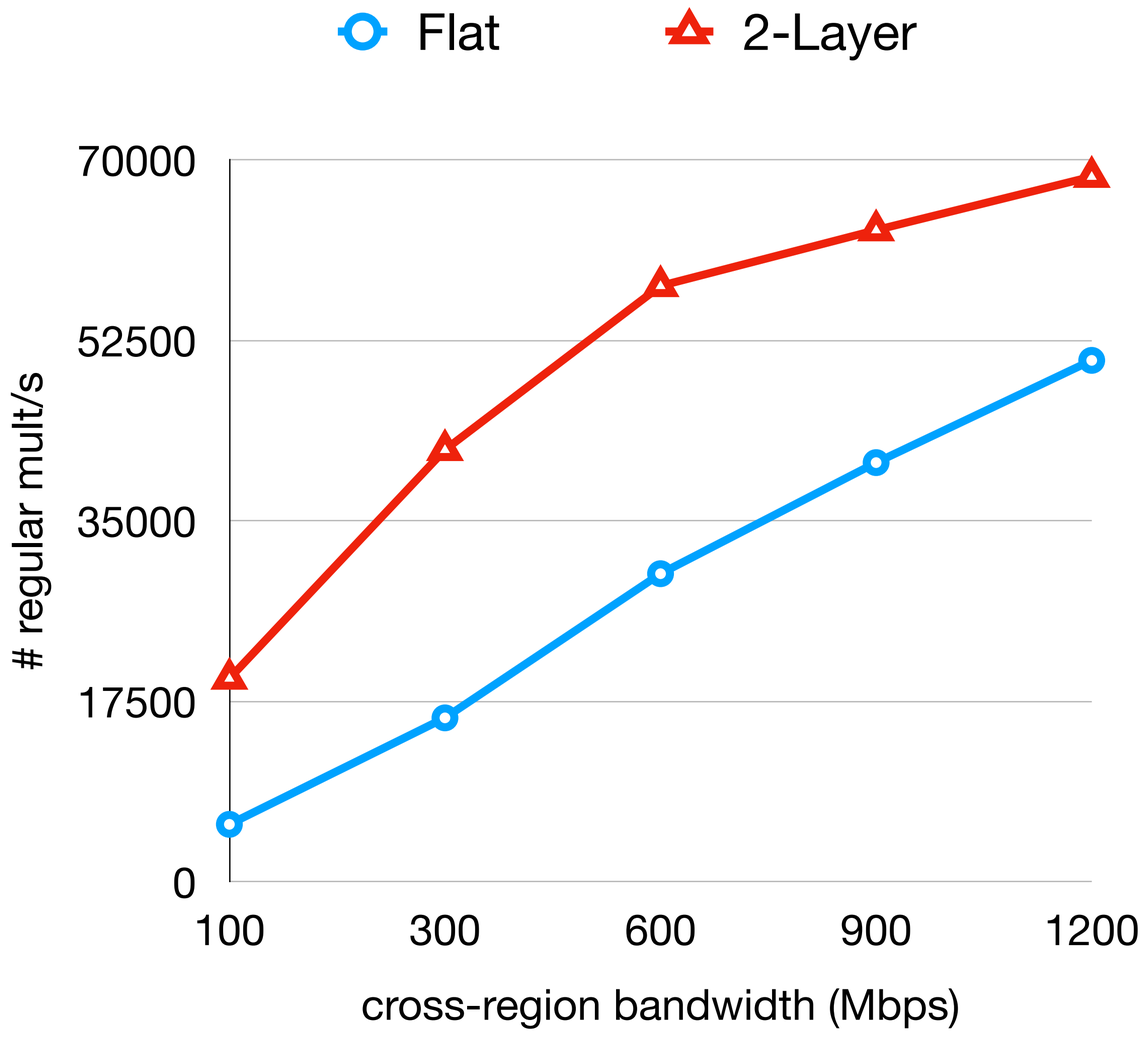
Impact of Cerebro's Physical Layout Optimizations



Impact of Cerebro's Physical Layout Optimizations



Impact of Cerebro's Physical Layout Optimizations



Up to 3.5x performance improvement

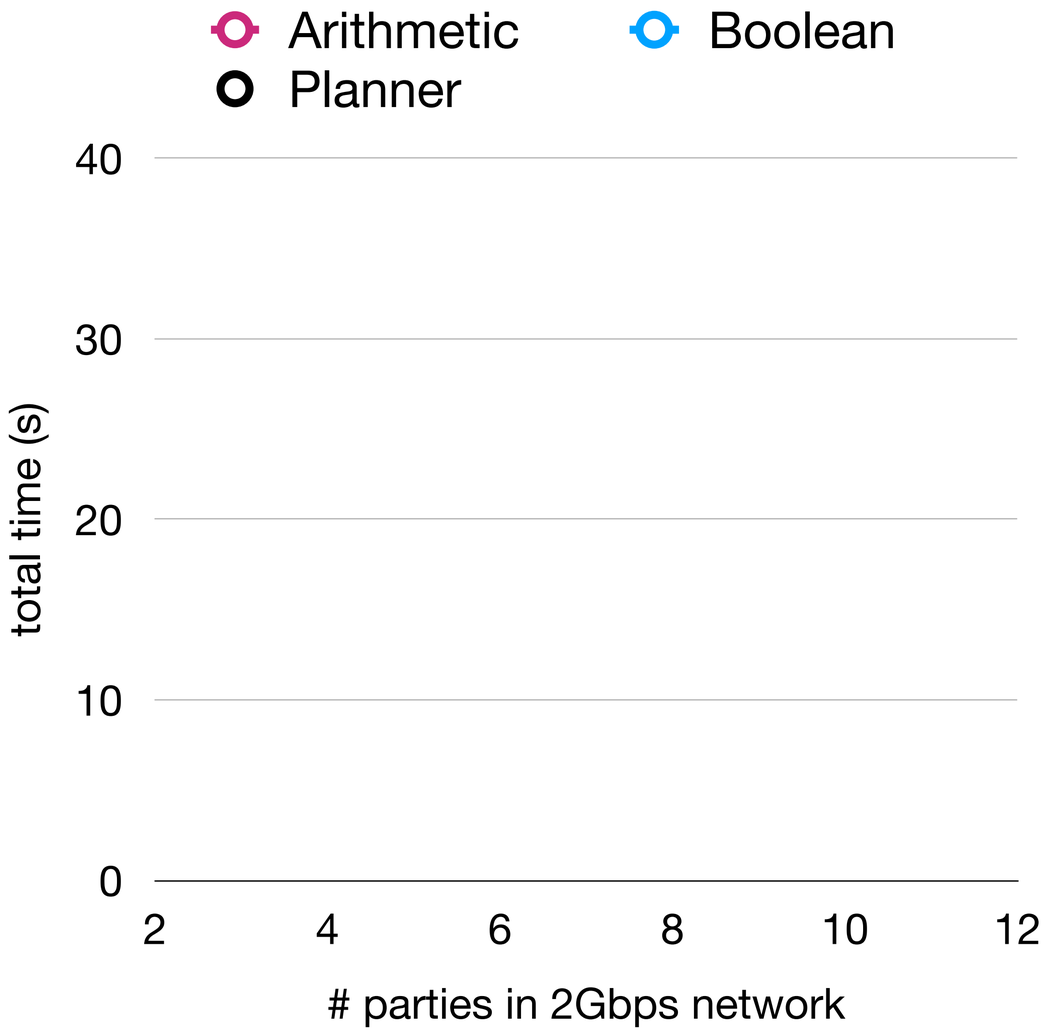
Effectiveness of Planner Decision Making

Effectiveness of Planner Decision Making

- Performance improvement of Cerebro's physical planning when compared to the worst case *secure* plan generated without Cerebro.

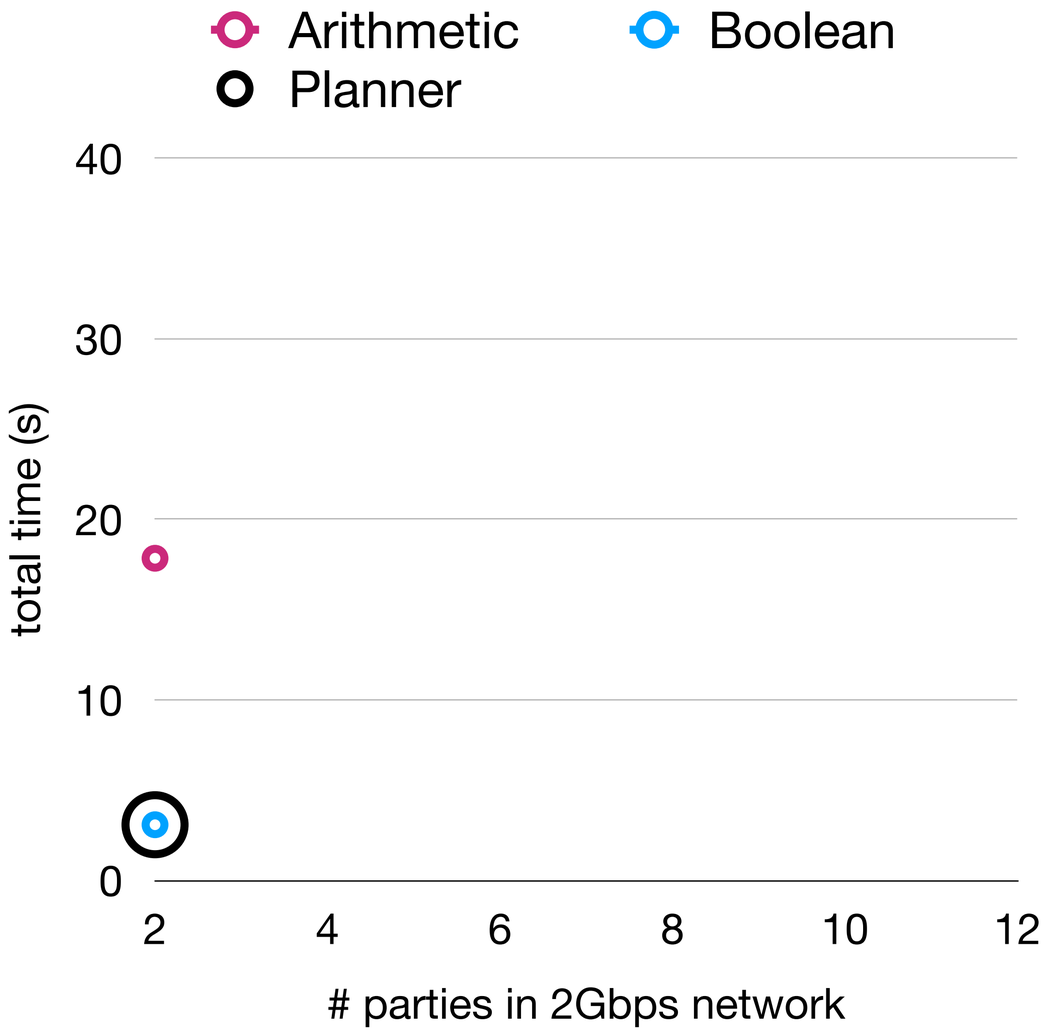
Effectiveness of Planner Decision Making

Effectiveness of Planner Decision Making



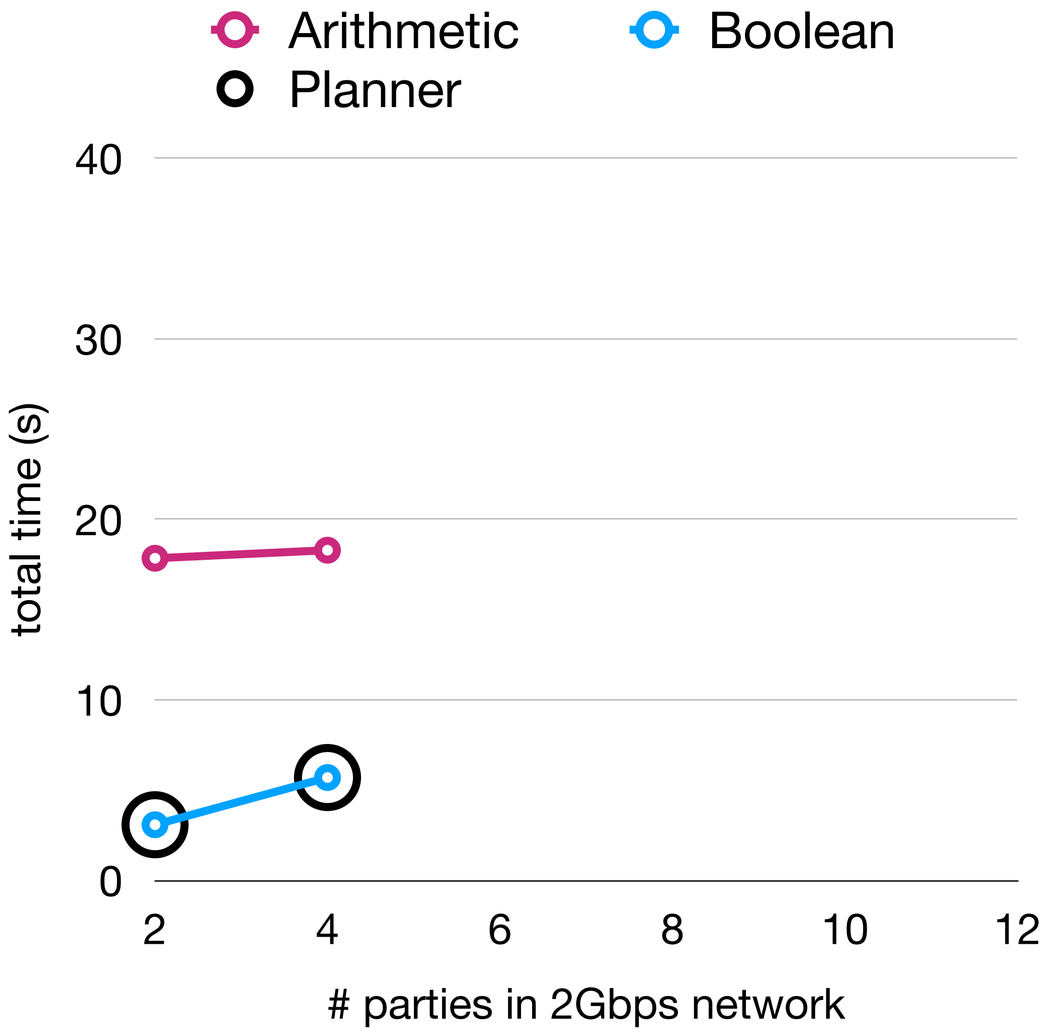
10-layer decision tree inference

Effectiveness of Planner Decision Making



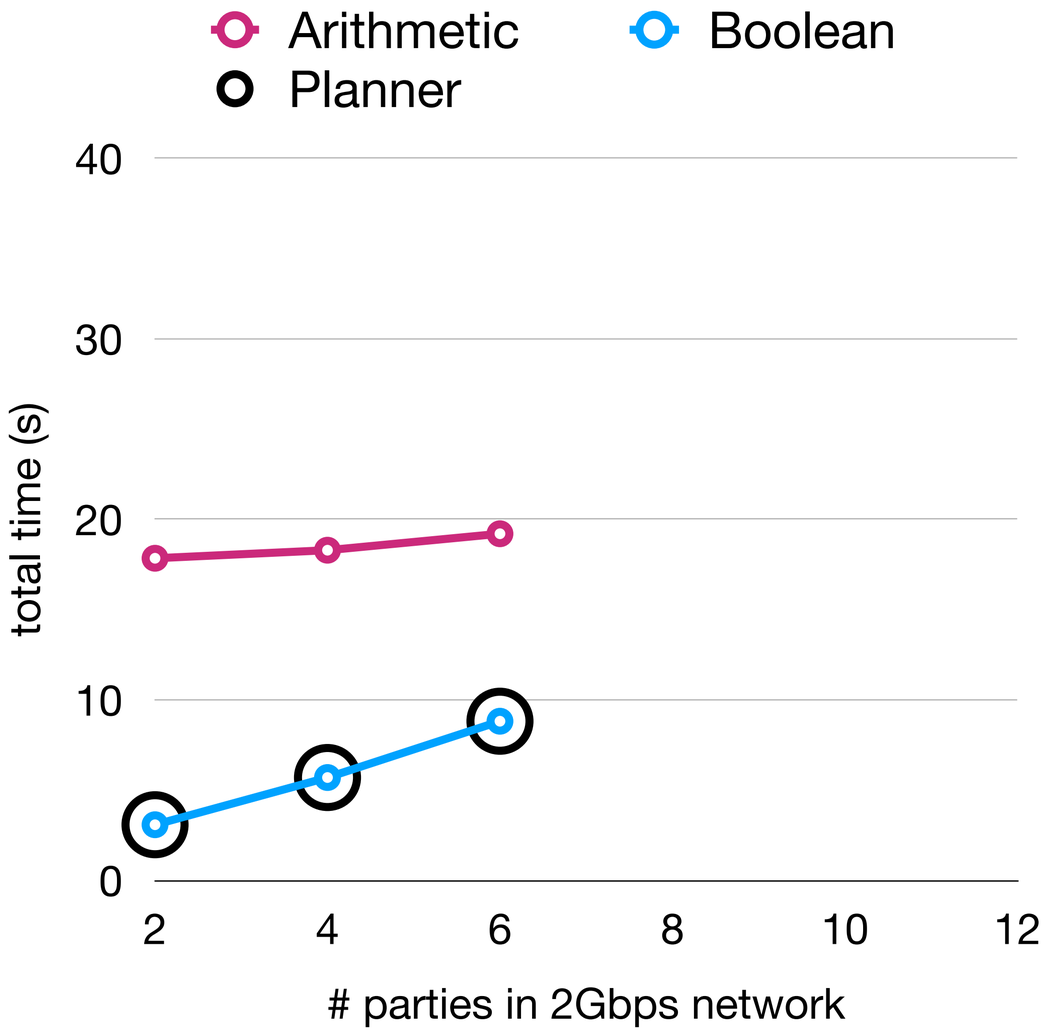
10-layer decision tree inference

Effectiveness of Planner Decision Making



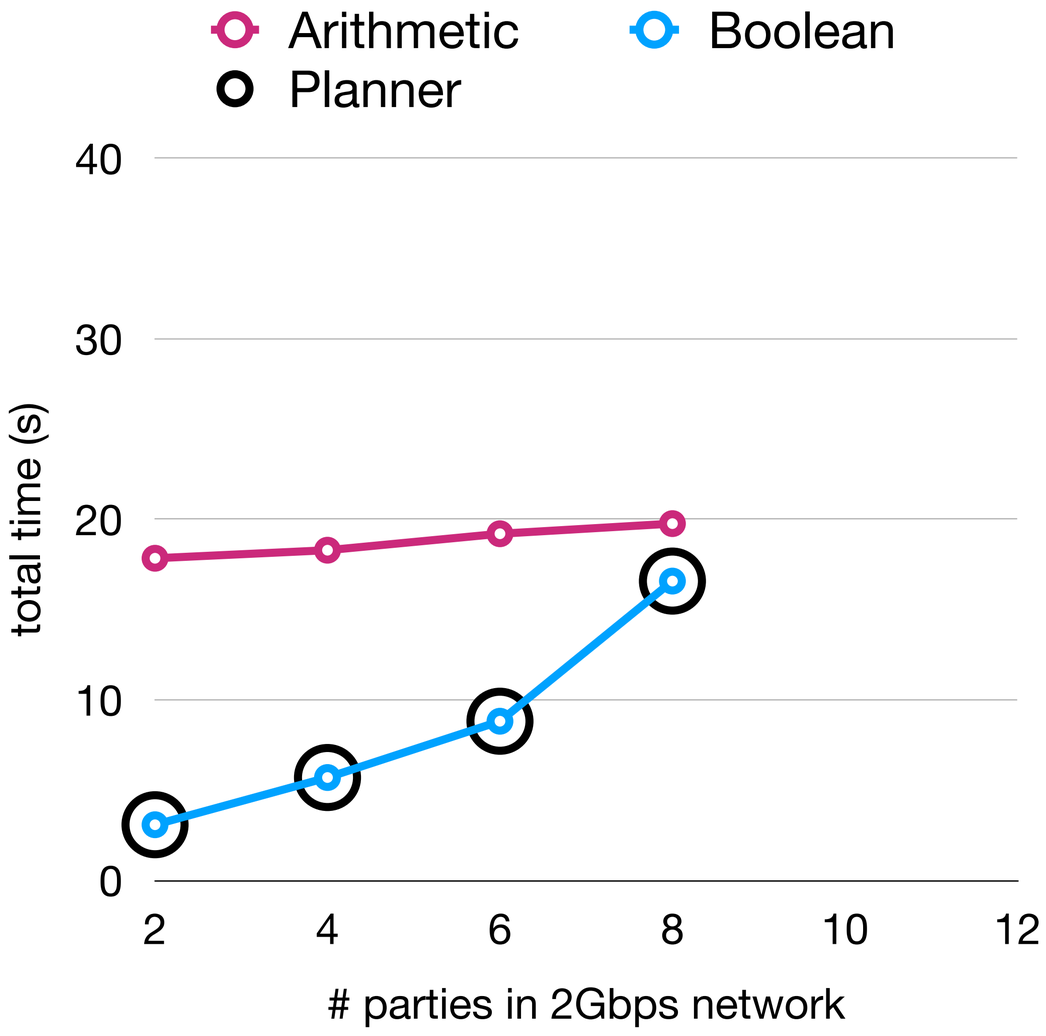
10-layer decision tree inference

Effectiveness of Planner Decision Making



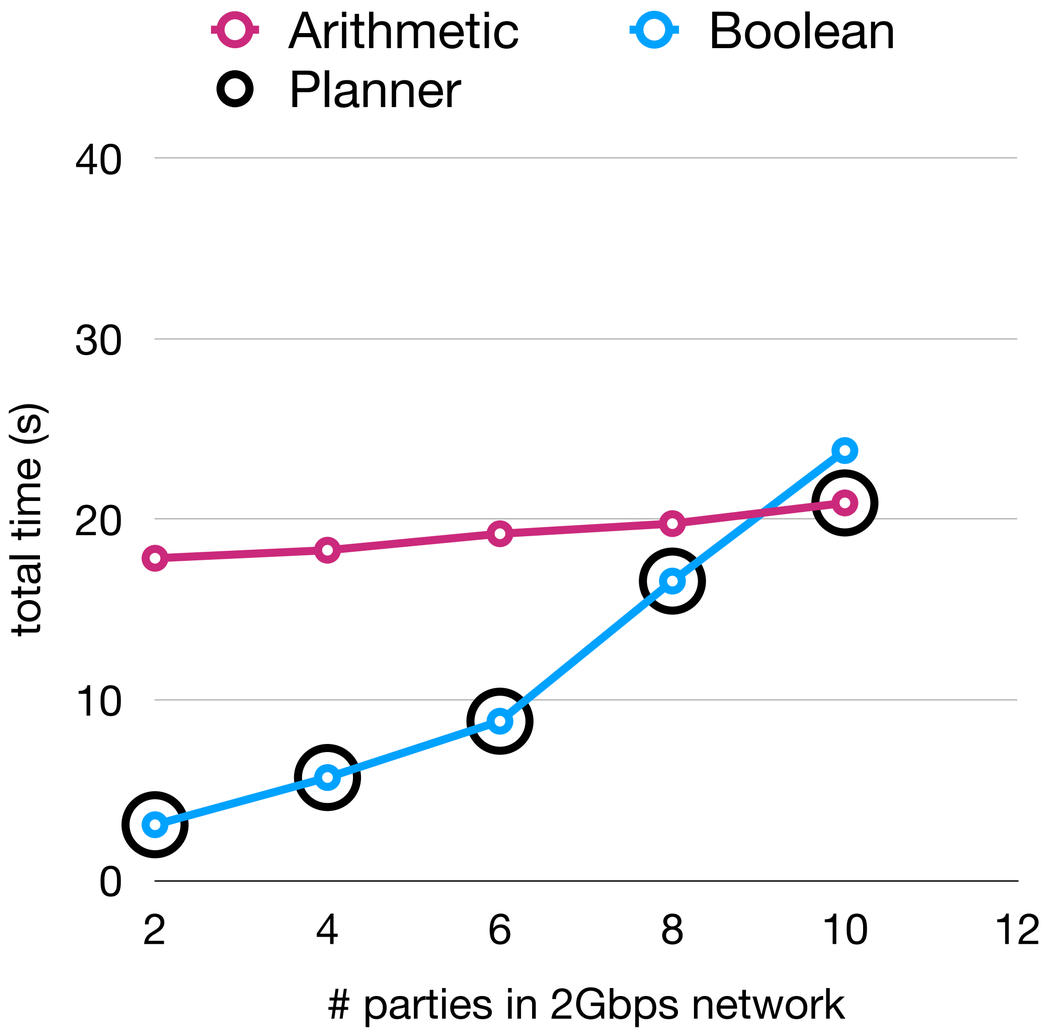
10-layer decision tree inference

Effectiveness of Planner Decision Making



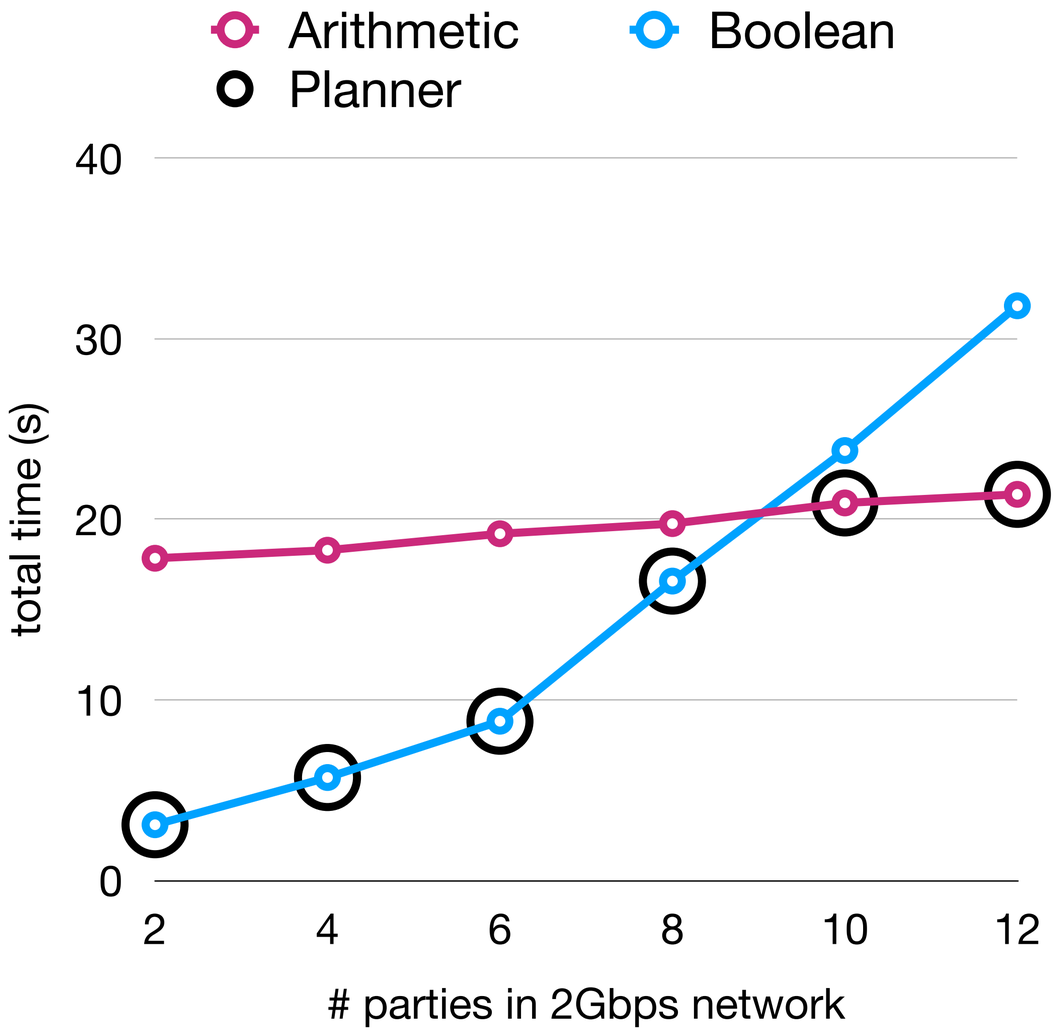
10-layer decision tree inference

Effectiveness of Planner Decision Making



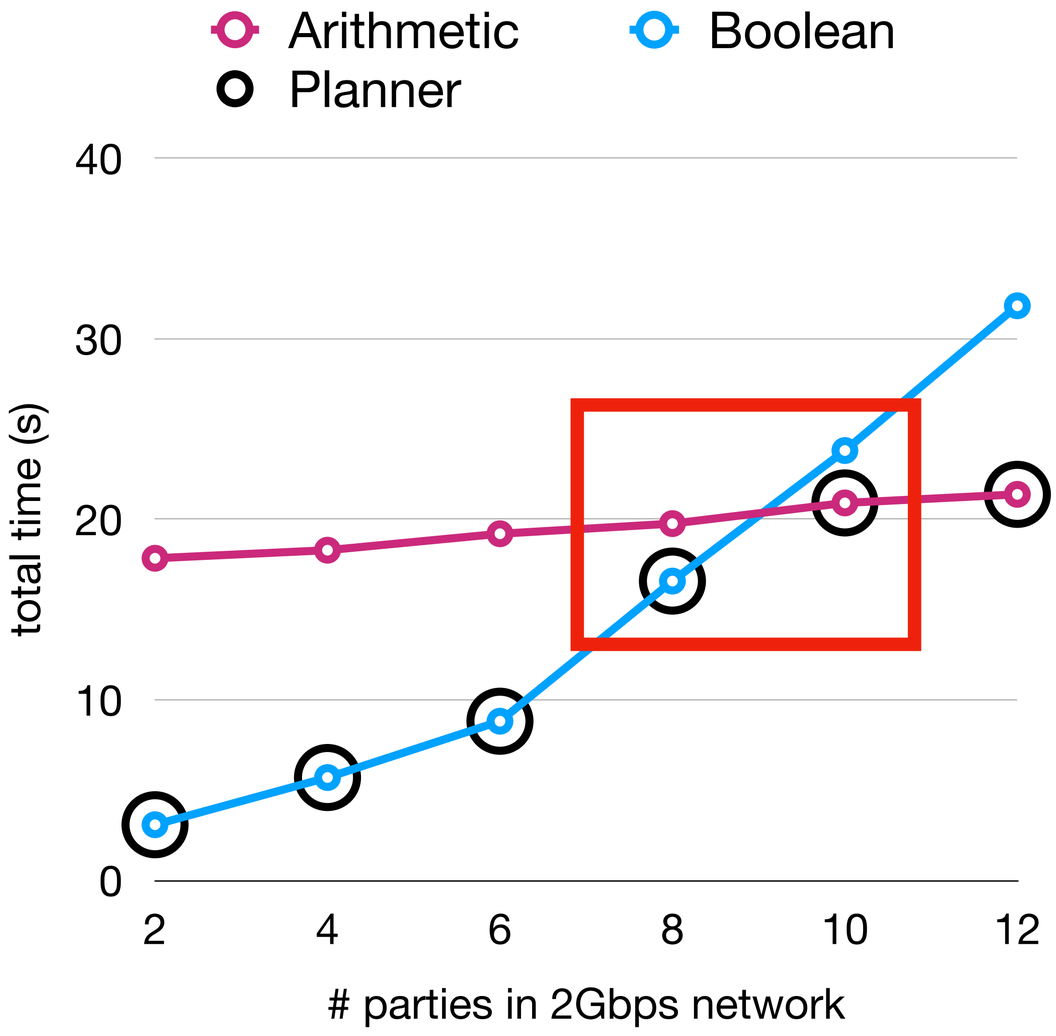
10-layer decision tree inference

Effectiveness of Planner Decision Making



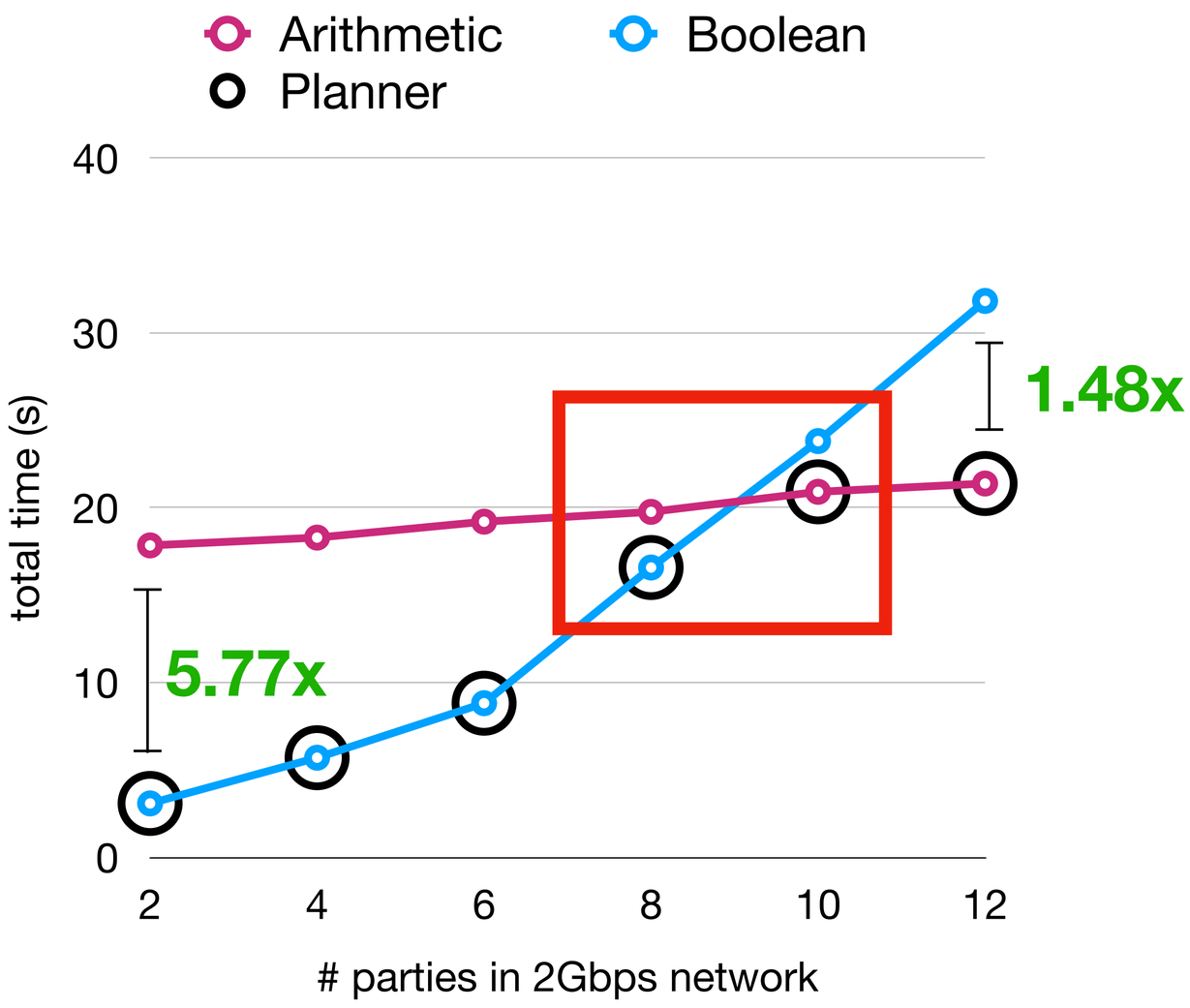
10-layer decision tree inference

Effectiveness of Planner Decision Making



10-layer decision tree inference

Effectiveness of Planner Decision Making



10-layer decision tree inference

Conclusion

Conclusion

- Cerebro is a programmable secure collaborative learning platform.

Conclusion

- Cerebro is a programmable secure collaborative learning platform.
- Addresses the challenges that arise when deploying secure multiparty computation in practice.

Conclusion

- Cerebro is a programmable secure collaborative learning platform.
- Addresses the challenges that arise when deploying secure multiparty computation in practice.
- Open source: https://github.com/mc2-project/cerebro

Conclusion

- Cerebro is a programmable secure collaborative learning platform.
- Addresses the challenges that arise when deploying secure multiparty computation in practice.
- Open source: https://github.com/mc2-project/cerebro
- Contact: ryan.deng@berkeley.edu