CHINA · SHENZHEN

Data Outsourcing in Cloud Computing: Reliability, Security and Privacy

Ning Cao
WalmartLabs Engineering Manager

CNUTCon 2017
全球运维技术大会

上海·光大会展中心大酒店 | 2017.9.10–11

智能时代的新运维

大数据运维
DevOps
安全
SRE
Kubernetes
Serverless
游戏运维
AIOps
智能化运维
基础架构
监控
互联网金融

主办方 Geekbang. 极客邦科技 InfoQ

StuQ

斯达克学院

实践驱动的IT教育

http://www.stuq.org

斯达克学院(StuQ)， 极客邦旗下实践驱动的IT教育平台。通过线下和线上多种形式的综合学习解决方案，帮助IT从业者和研发团队提升技能水平。

人工智能

大数据

前端开发

后端开发

架构设计

移动开发

运维设计

产品测试

产品经理

技术管理

10大职业技术领域课程

SPEAKER
INTRODUCE

曹宁 WalmartLabs Engineering Manager

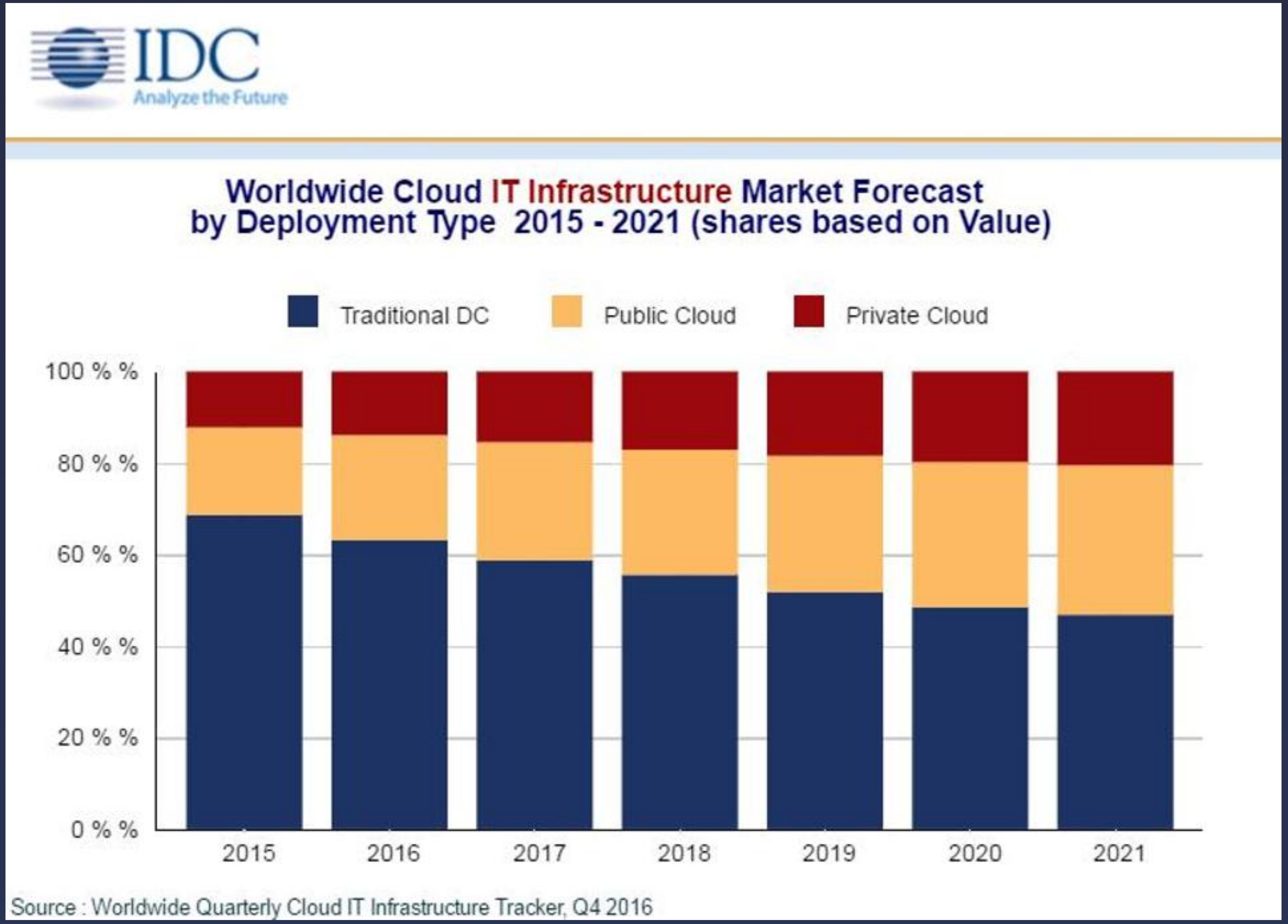
- Ning Cao is an engineering manager in search runtime team at WalmartLabs. Prior to that, he worked at Google, Huawei.
- Ning received his Ph.D. in Electrical and Computer Engineering at Worcester Polytechnic Institute. His publications have 4000+ citations.

TABLE OF

CONTENTS 大纲

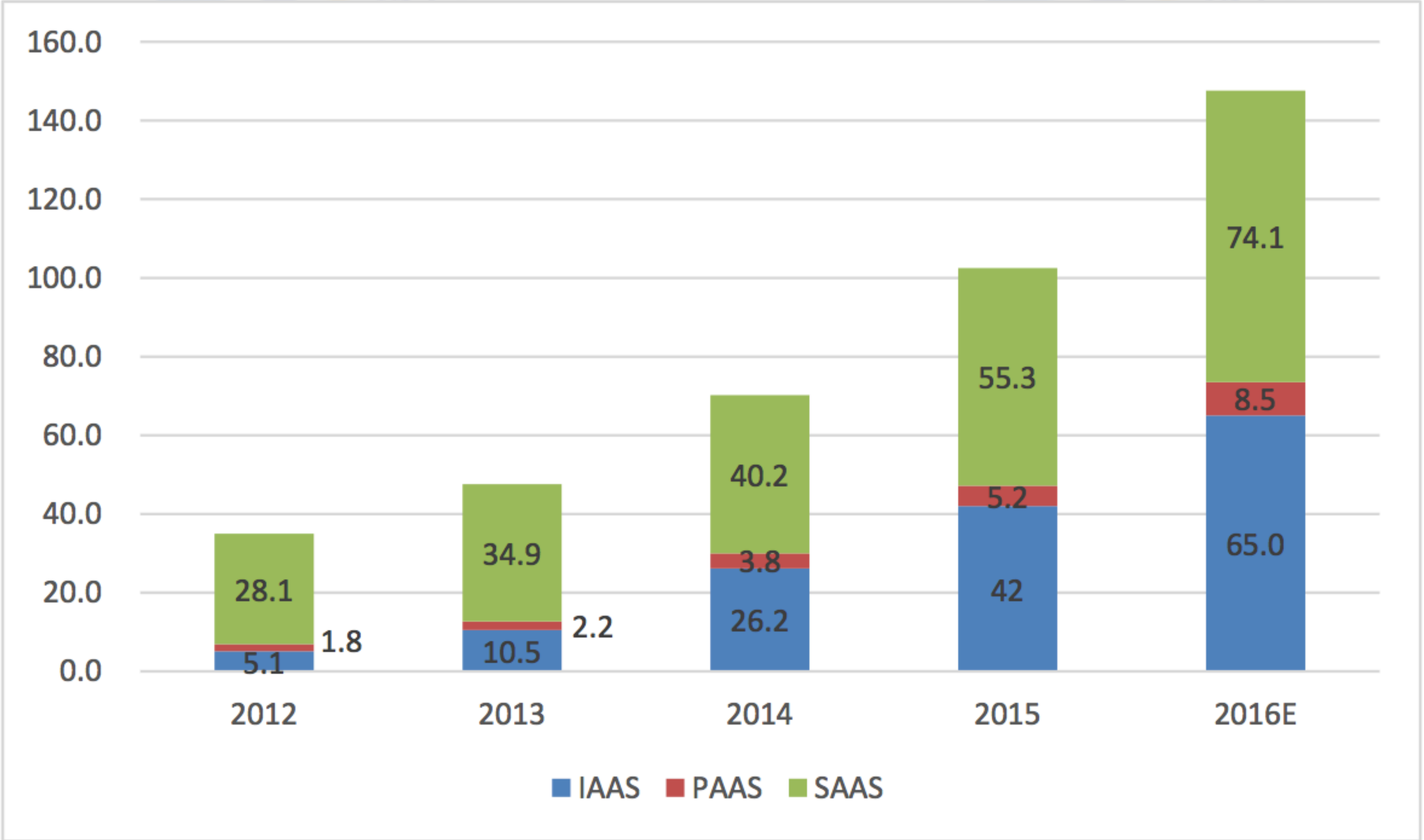
Data Outsourcing in Cloud Computing

- Cloud Computing
  - great flexibility
  - economic savings



Data Outsourcing in Cloud Computing

- Public Cloud in China
  - IAAS: fast growing
  - 70%: cloud host, cloud storag



数据来源：中国信通院

图 9　公共云细分市场规模（单位：亿元人民币）

Data Outsourcing in Cloud Computing

- Cloud Customers
  - Current: internet companies
    - Game, e-commerce, mobile, social, etc
  - Next/Ongoing: traditional industries
    - Government, finance/bank, health/medical/hospital, manufacturing, transport, etc.

Data Outsourcing in Cloud Computing

- Sensitive data outsourcing in public/hybrid Cloud
    - Data owner: government, finance/bank, health/medical/hospital, etc.
    - Requirement: data ownership, responsibility
    - Concerns: reliability, availability, security, privacy, integrity, etc.

TABLE OF

CONTENTS 大纲

Unreliability of Cloud Storage

- Byzantine failures
  - hardware errors
  - cloud maintenance personnel's misbehaviors
- External attacks
  - natural disasters, like fire and earthquake
  - malicious hacking, e.g., pollution attack, or replay attack

Reliable Data Outsourcing

- How to ensure data reliability?
  - Adding data redundancy to multiple servers

TABLE OF

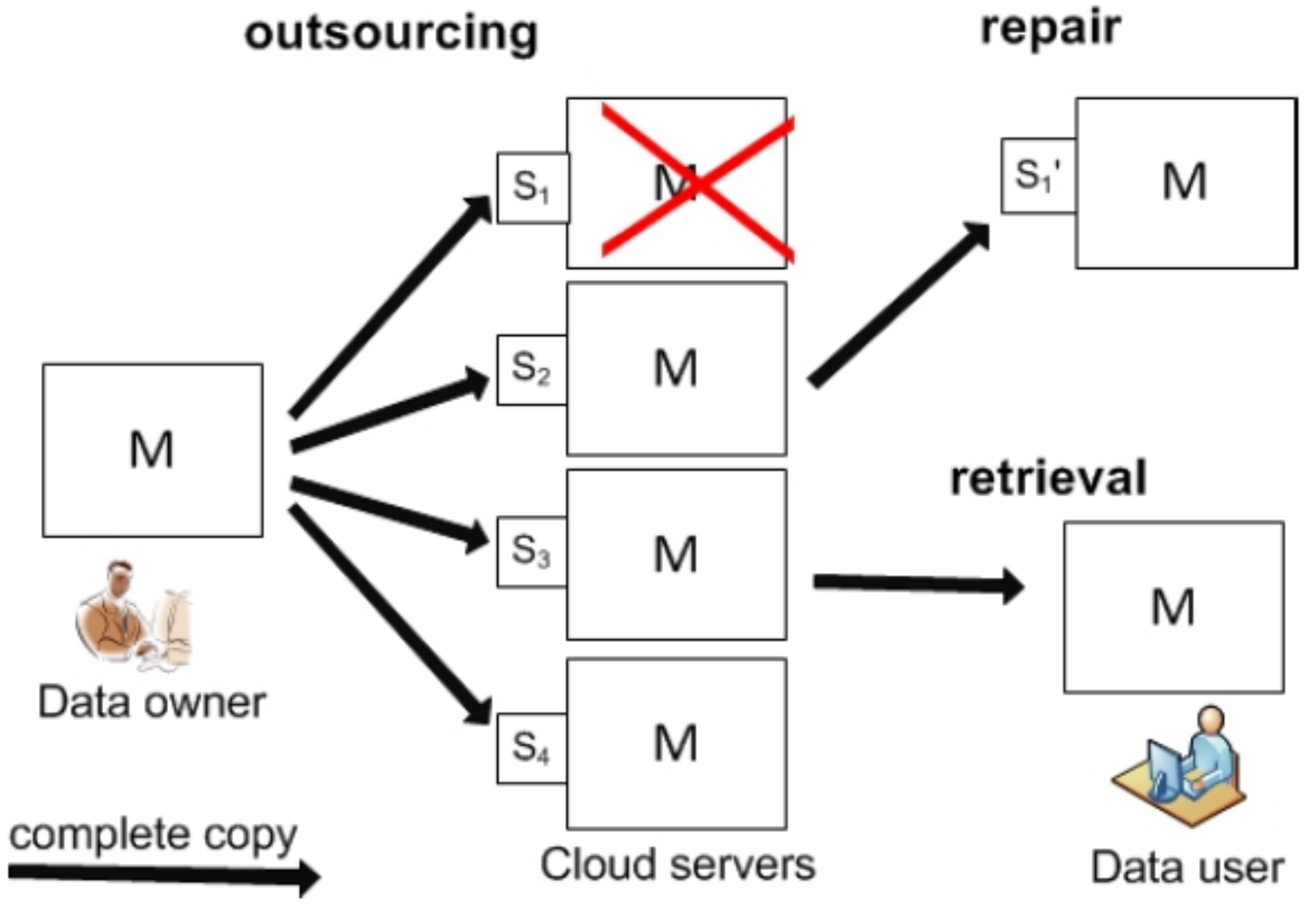
CONTENTS 大纲

- Data Outsourcing in Cloud Computing
- Reliable Data Outsourcing
    - Redundancy Techniques
    - Fountain Codes Based Reliable Storage
- Search over Encrypted Cloud Data

Redundancy Techniques

- Replication-based
  - Pros: simple data management
  - Cons: high storage cost

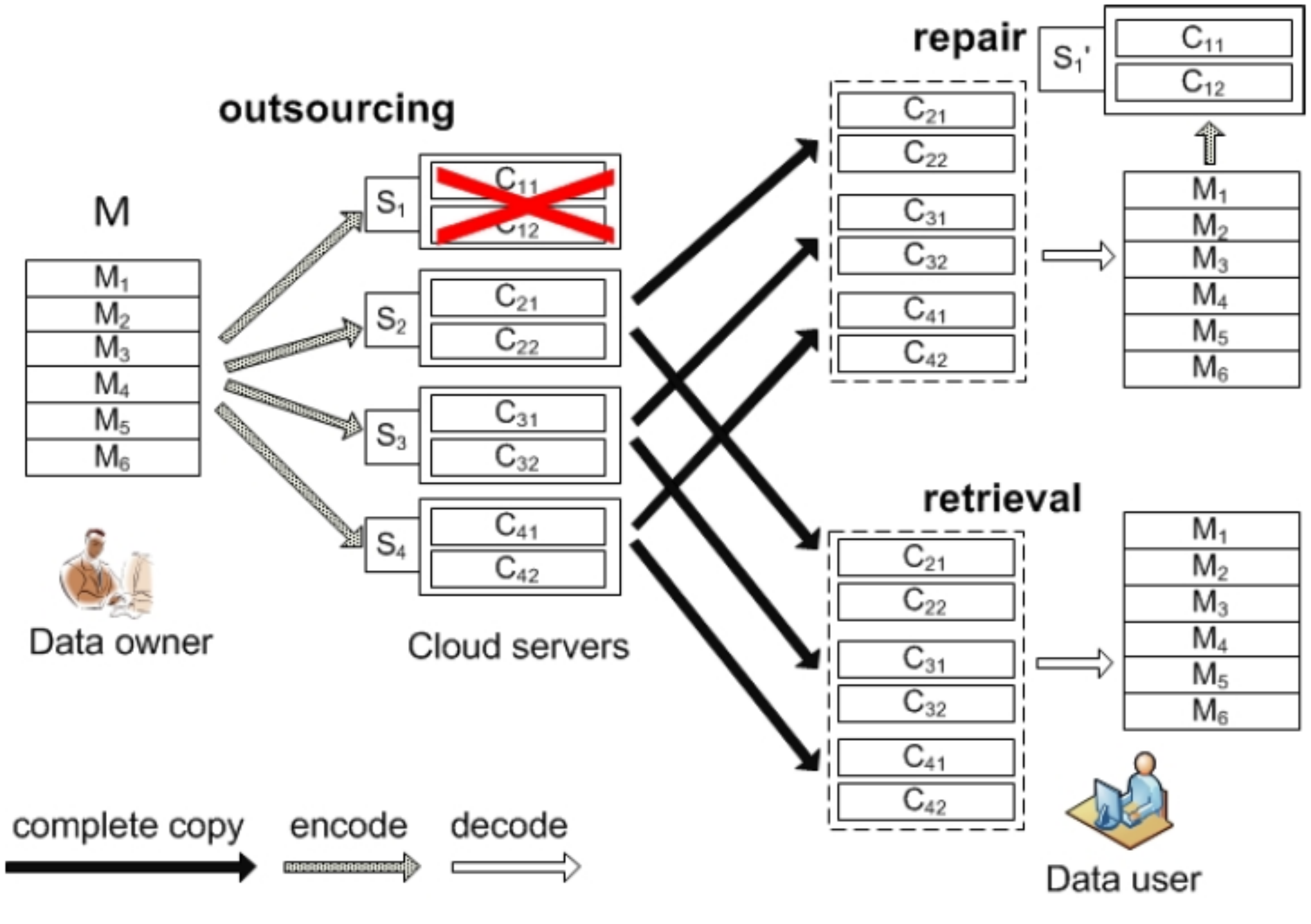
    low throughput



(a) Replication

Redundancy Techniques

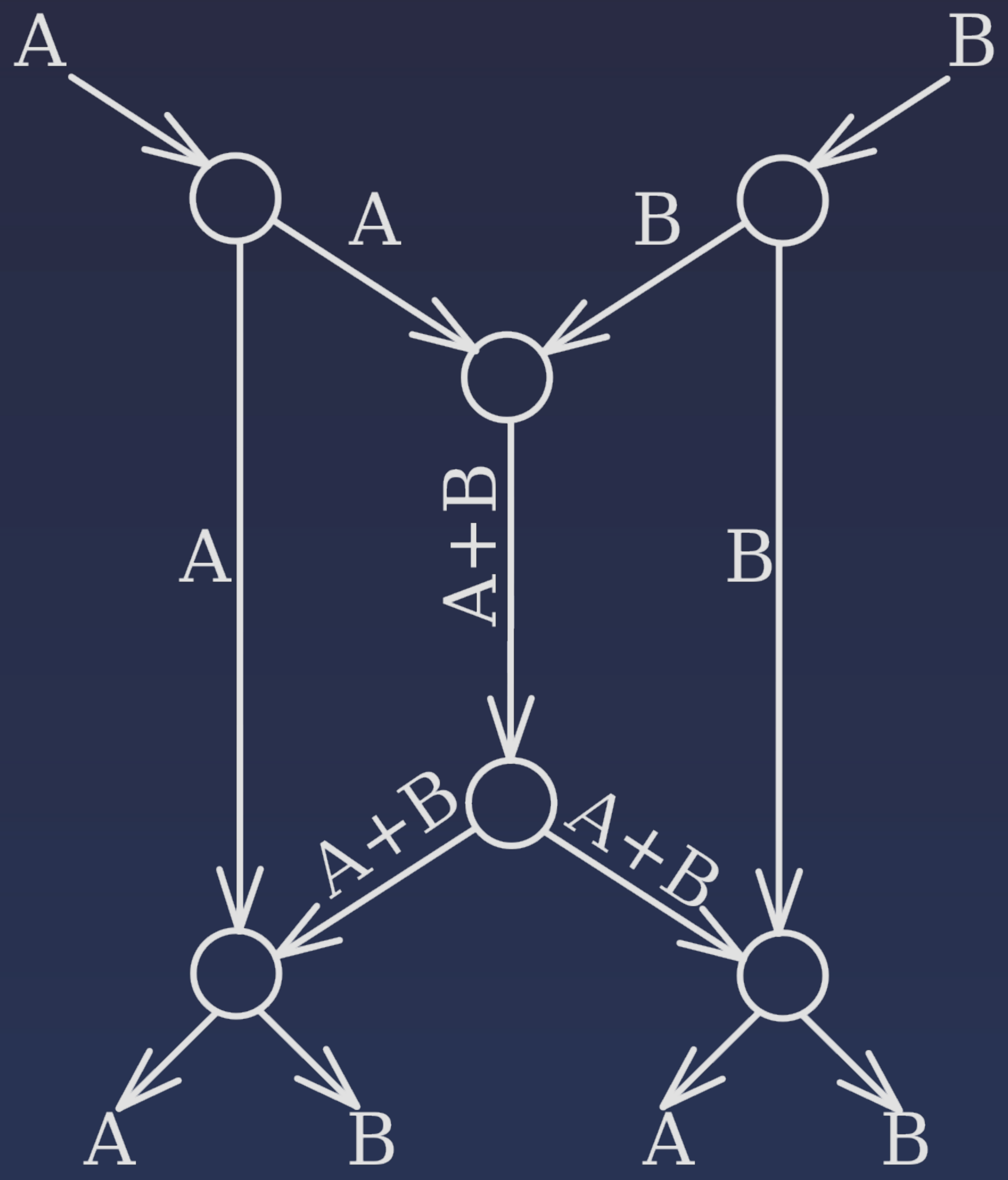
- Erasure codes-based
  - Pros: much less data redundancy
    high throughput
  - Cons: less repair communication



(b) Optimal Erasure codes

Redundancy Techniques

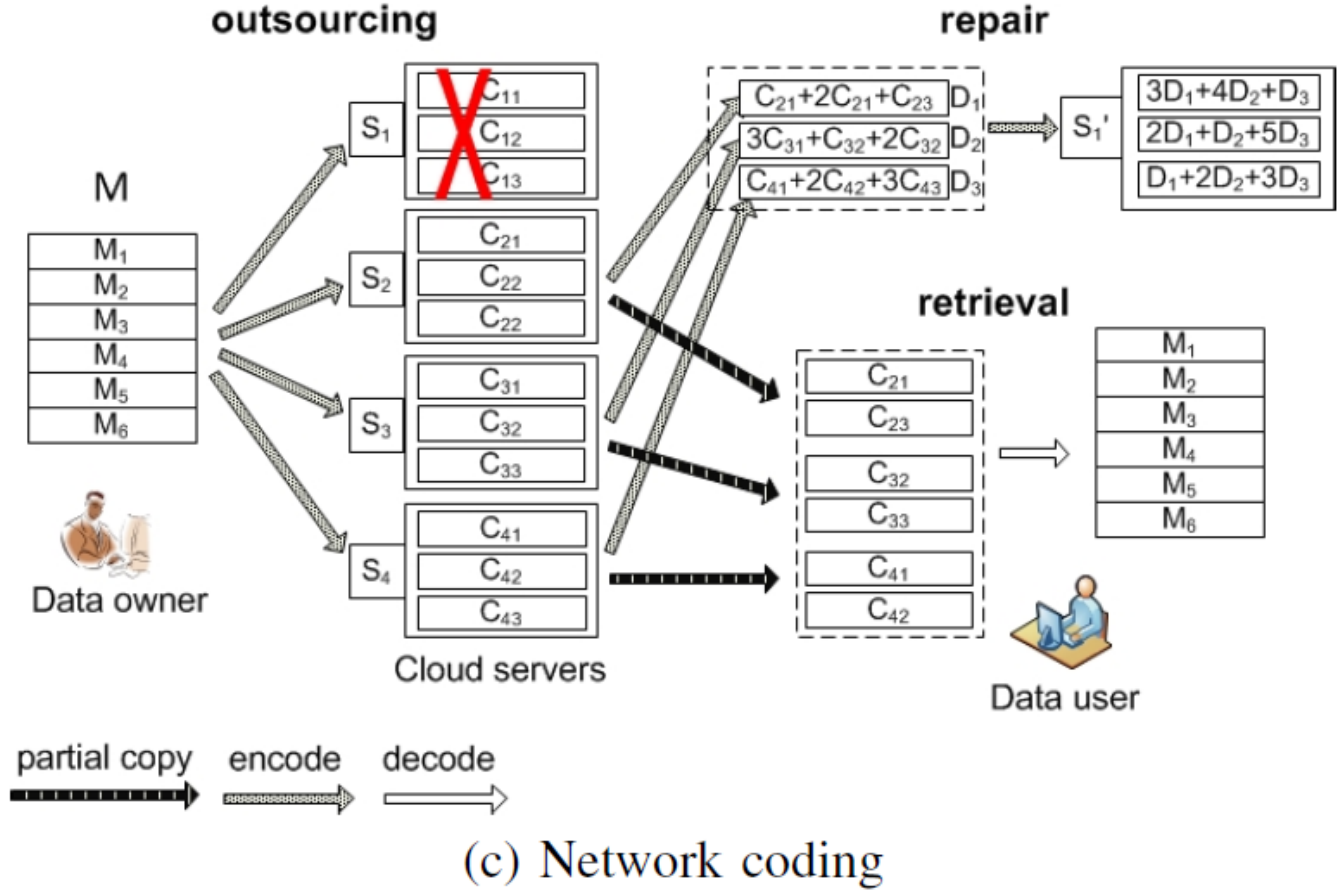
- Network coding-based
  - networking technique
  - increase network throughput



Redundancy Techniques

- Network coding-based
  - Pros: less repair communication
  - Cons: high decoding cost

    availability

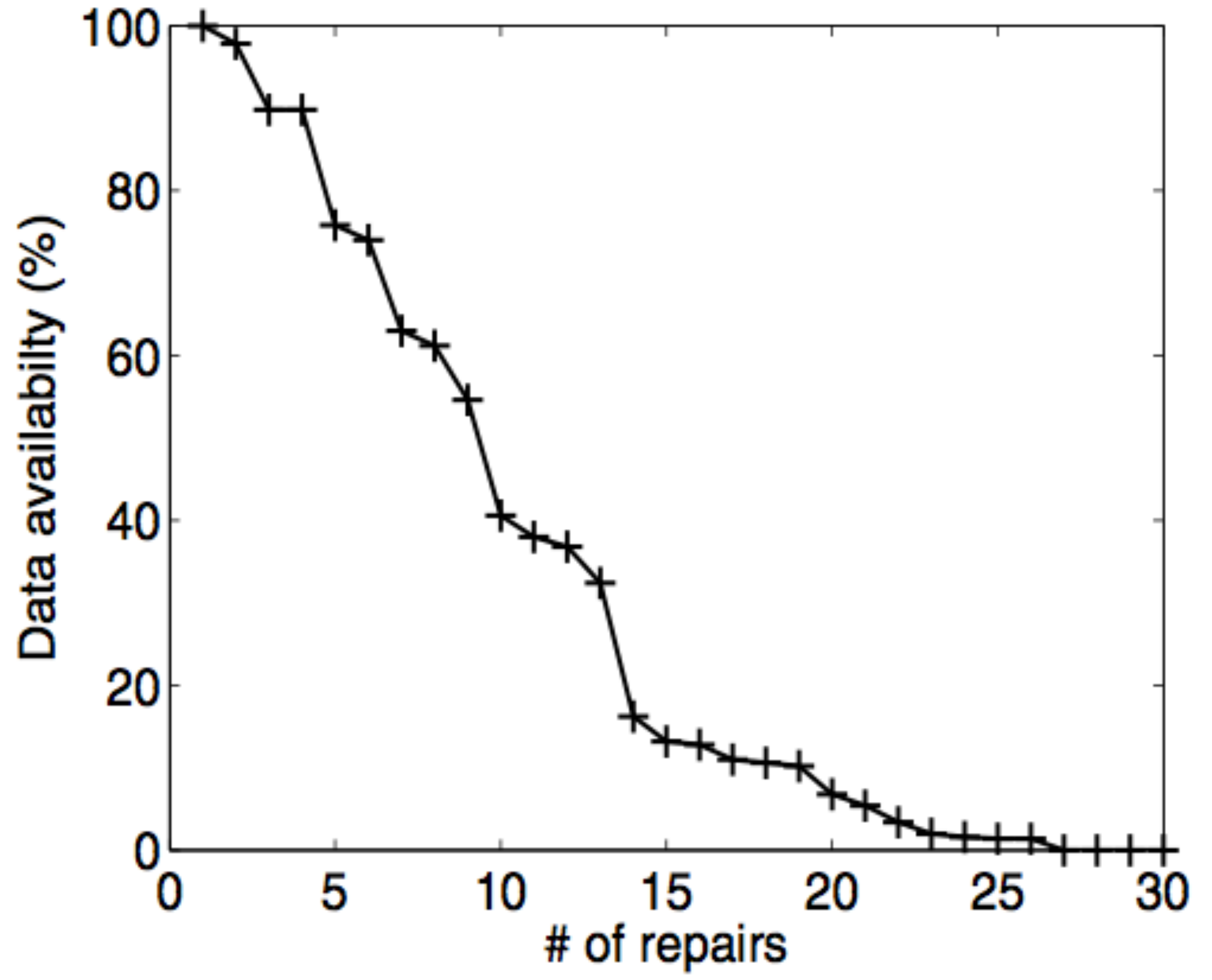


(c) Network coding

Redundancy Techniques

- Network coding-based
  - Pros: less repair communication
  - Cons: high decoding cost

decreasing availability after rep



Data availability after functional repair as in LTNC.

Data Reliability

- How to perform data repair and data retrieval at minimal cost in cloud?
  - Both data storage and transmission are charged
    - “pay-as-you-use”
    - low storage, computation and communication cost

TABLE OF

CONTENTS 大纲

Fountain Codes

- LT code (Luby transform code)
  - File M is split into m original packets, $M_1, \ldots, M_m$
  - Generate $n\alpha$ encoded packets following LT codes (bitwise XOR)
  - α is the number of packets outsourced to each storage server
    - $\alpha = m/k(1+\varepsilon)$
  - Any k servers have totally $m(1+\varepsilon)$ encoded packets

Fountain Codes

- LT code (Luby transform code)
  - Near-optimal erasure codes
    - all m original packets can be recovered from any $m(1+\epsilon)$ encoded packets with probability $1-\delta$
  - Efficient decoding $O(m*\ln m)$: Fast Belief Propagation decoder
- Challenges to utilize LT code: Decodability; Efficient data repair

Data Decodability

- How to satisfy the data availability requirement in cloud storage ?
  - Goal: all m original packets can be recovered from any $m(1+\epsilon)$ encoded packets with probability 100% (vs. $1-\delta$)
    - Divide all the encoded packets equally into n groups
    - Run the Belief Propagation decoder on every k-combination of n groups
    - If decoding fails, regenerate encoded packets until successful

Data Repair

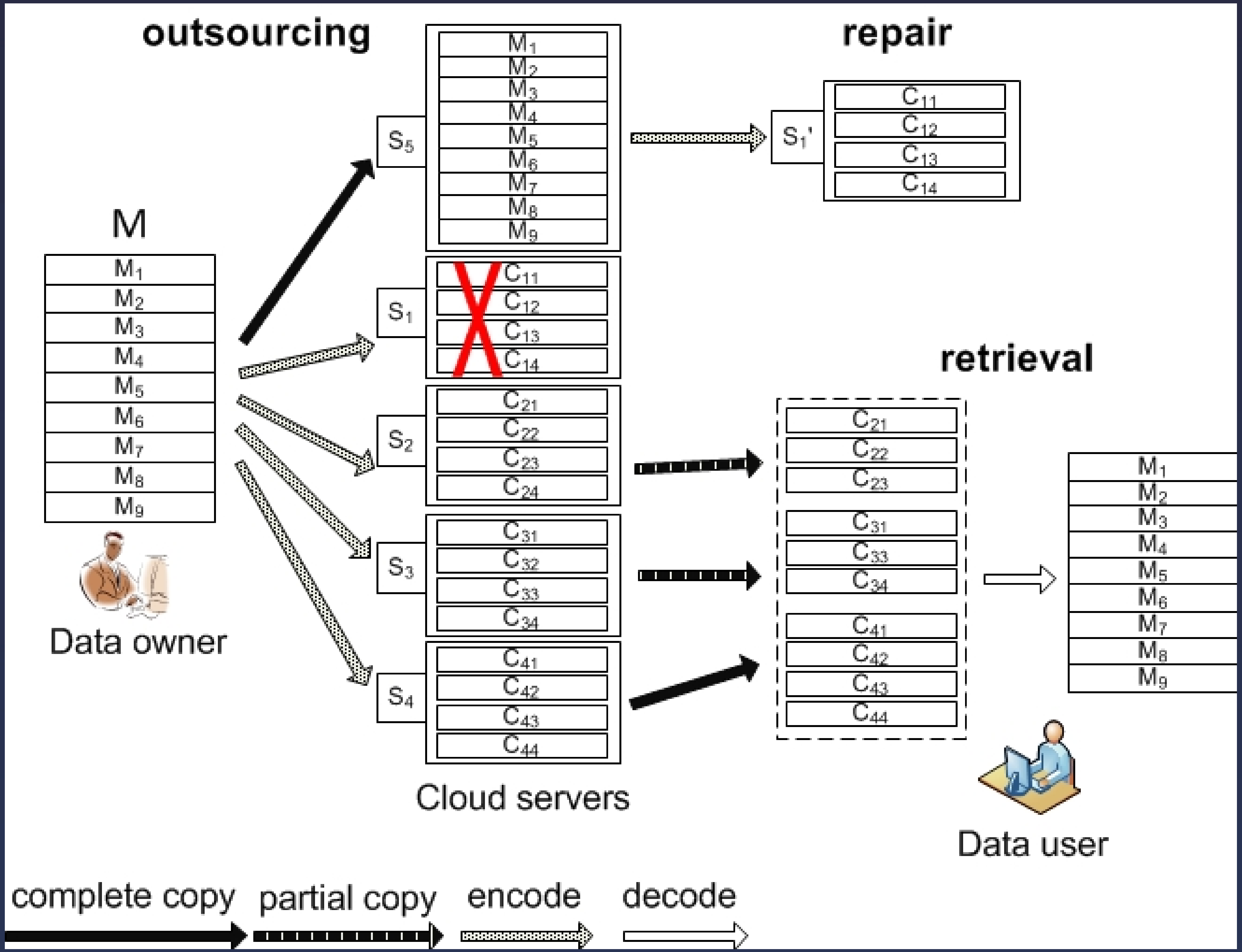
- Exact repair
  - generate exactly same packets as those previously stored in corrupted servers
  - do not introduce linear dependence: maintain the data availability
- Functional repair
  - generates correct encoded packets, but not exactly same as those corrupted
  - random linear recoding cannot satisfy the degree requirement in LT codes

Data Repair

- How to do exact repair?
  - A straightforward data repair method
    - recover all original data packets if a storage server is corrupted
    - do the encoding to generate coded packets
  - Introduce much cost of both computation and communication!

Data Repair

- Exact repair
  - One repair server S_{n+1}

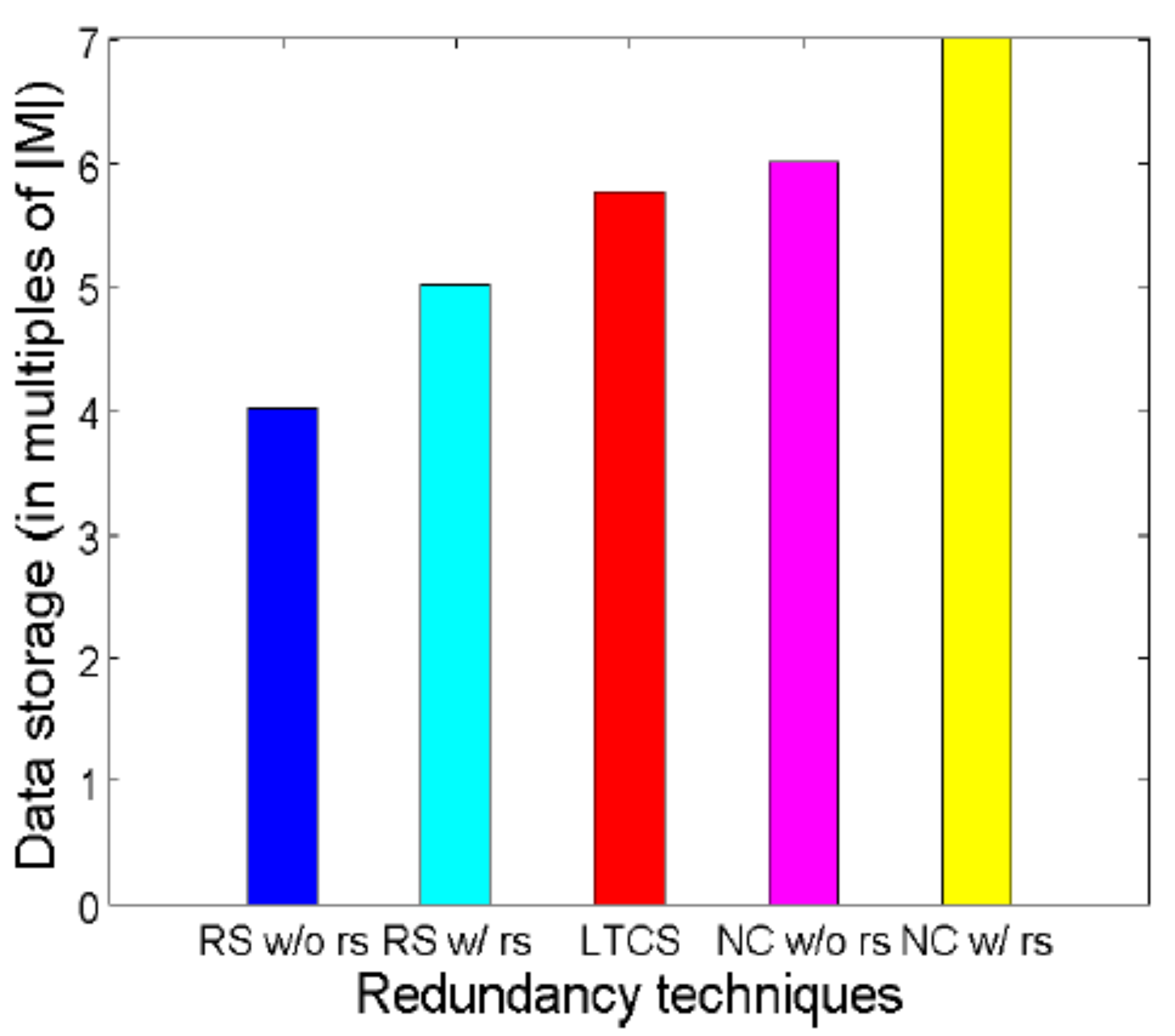


Complexity Analysis

- Theoretical complexity analysis (introduce repair server)

	Network Coding	Reed-Solomon	LTCS
Total server storage	$O((2n/(k+1)) \cdot \lvert\mathcal{M}\rvert)$	$O((1+n/k) \cdot \lvert\mathcal{M}\rvert)$	$O((1+n(1+\varepsilon)/k) \cdot \lvert\mathcal{M}\rvert)$
Encoding computation	$O(2nm^2/(k+1))$	$O(nm^2/k)$	$O((nm(1+\varepsilon)\ln m)/k)$
Retrieval communication	$O(\lvert\mathcal{M}\rvert)$	$O(\lvert\mathcal{M}\rvert)$	$O(\lvert\mathcal{M}\rvert)$
Retrieval computation	$O(m^2)$	$O(m^2)$	$\boxed{O(m \ln m)}$
Repair communication	$O(2T/(k+1) \cdot \lvert\mathcal{M}\rvert)$	$O(T(1/k+1/n) \cdot \lvert\mathcal{M}\rvert)$	$O(T((1+\varepsilon)/k+1/n) \cdot \lvert\mathcal{M}\rvert)$

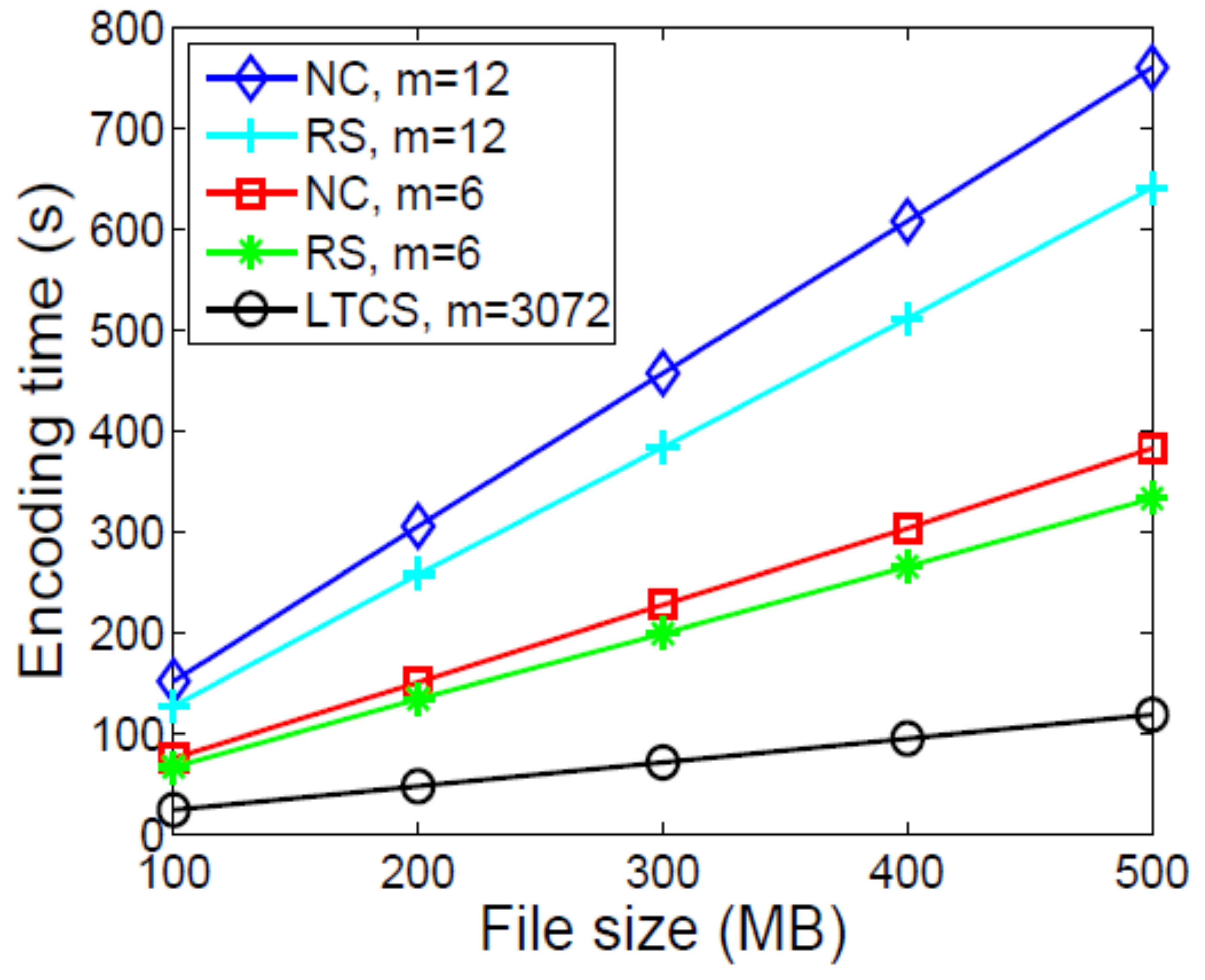
Experimental Evaluation



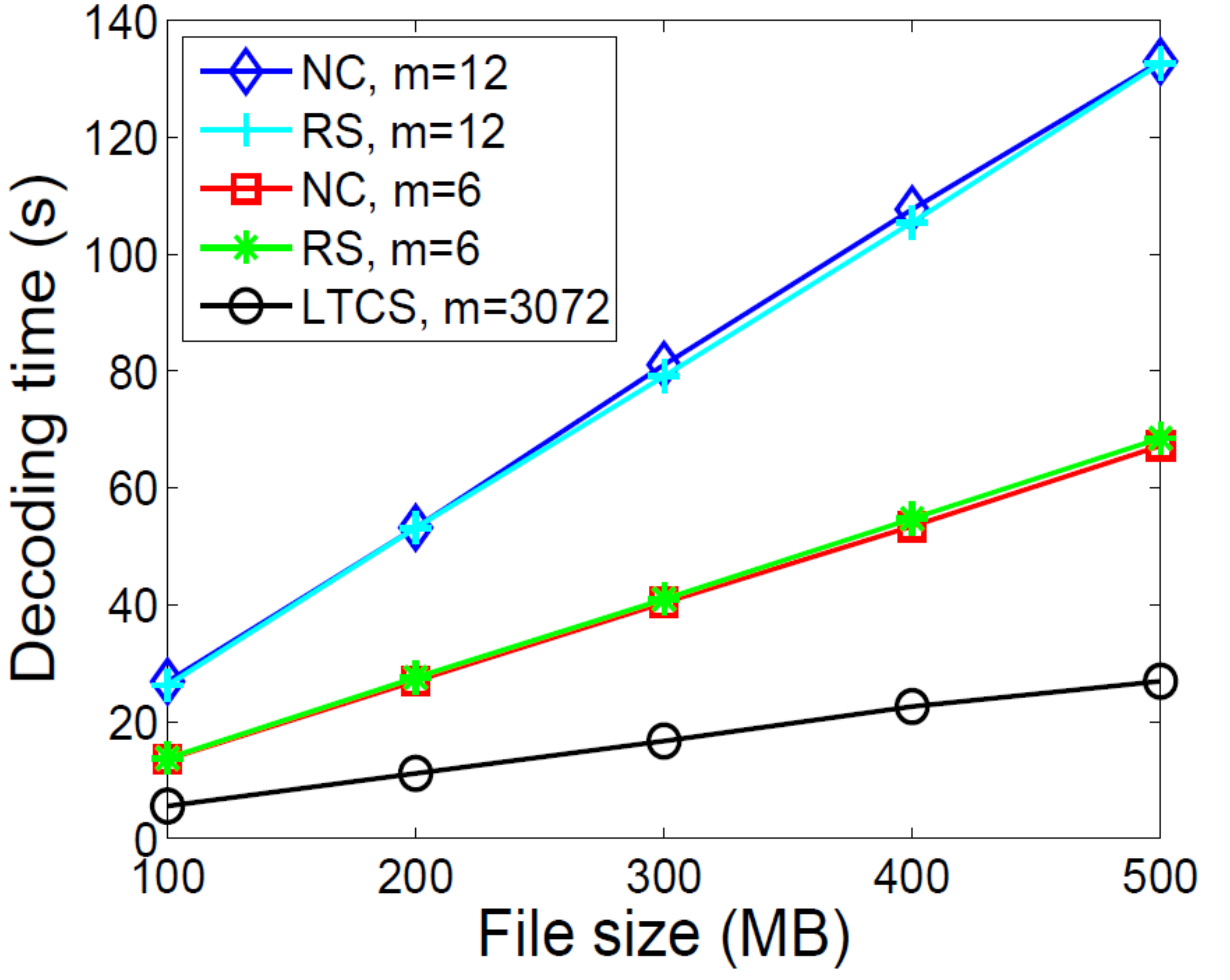
Data storage cost

主办方 Geekbang>. 极客邦科技 InfoQ

Experimental Evaluation

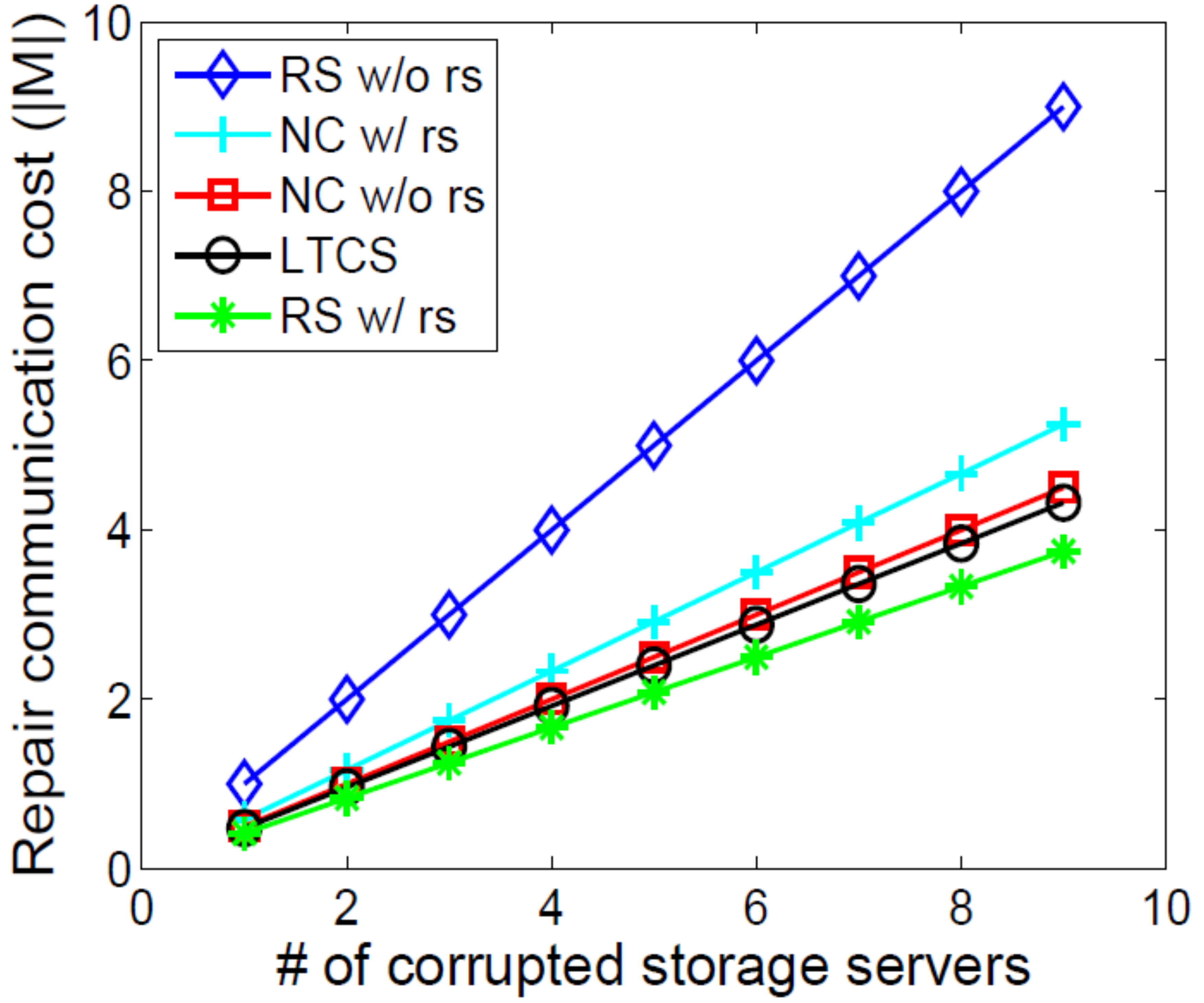


(a)Encoding



(b)Decoding

Experimental Evaluation



Communication cost of repair. n=12, k=3

TABLE OF
CONTENTS 大纲

- Data Outsourcing in Cloud Computing
- Reliable Data Outsourcing
- Search over Encrypted Cloud Data
    - Searchable Encryption
    - Predicate Encryption

Data Outsourcing in Cloud Computing

- Sensitive Data have to be encrypted before outsourcing
  - protect data privacy and combat unsolicited accesses
- Encryption makes data utilization a challenging task
  - traditional plaintext search -> no privacy guarantees
  - downloading all data and decrypting locally -> impractical

TABLE OF

CONTENTS 大纲

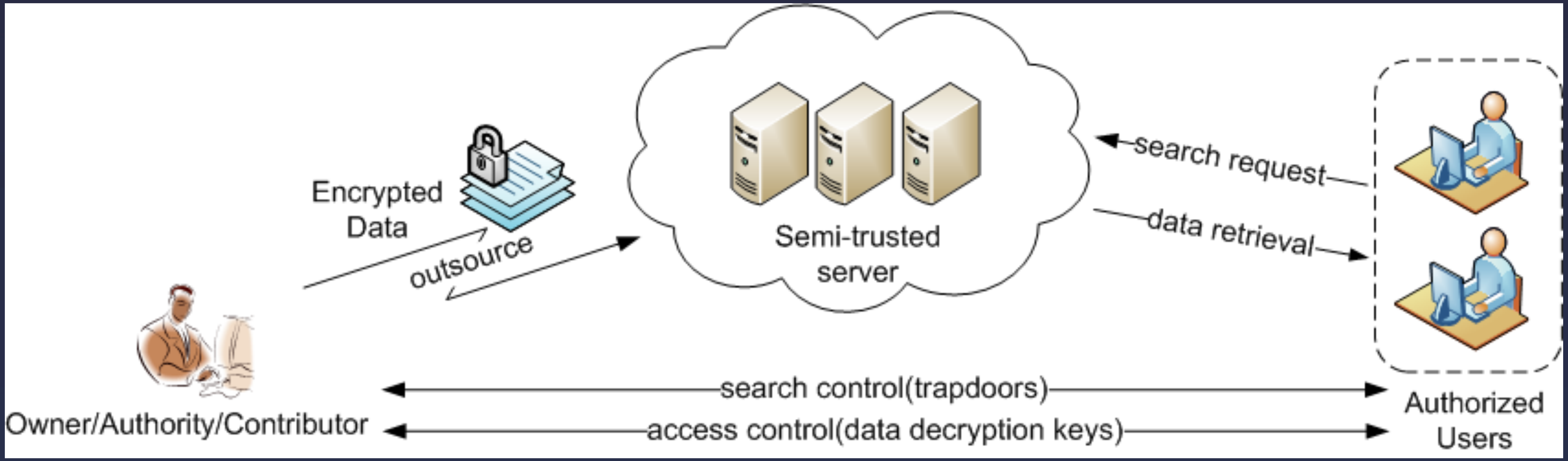
Searchable Encryption

- Two Categories based on the data contribution

Single Data Contributor

Multiple Data Contributor

Single Data Contributor

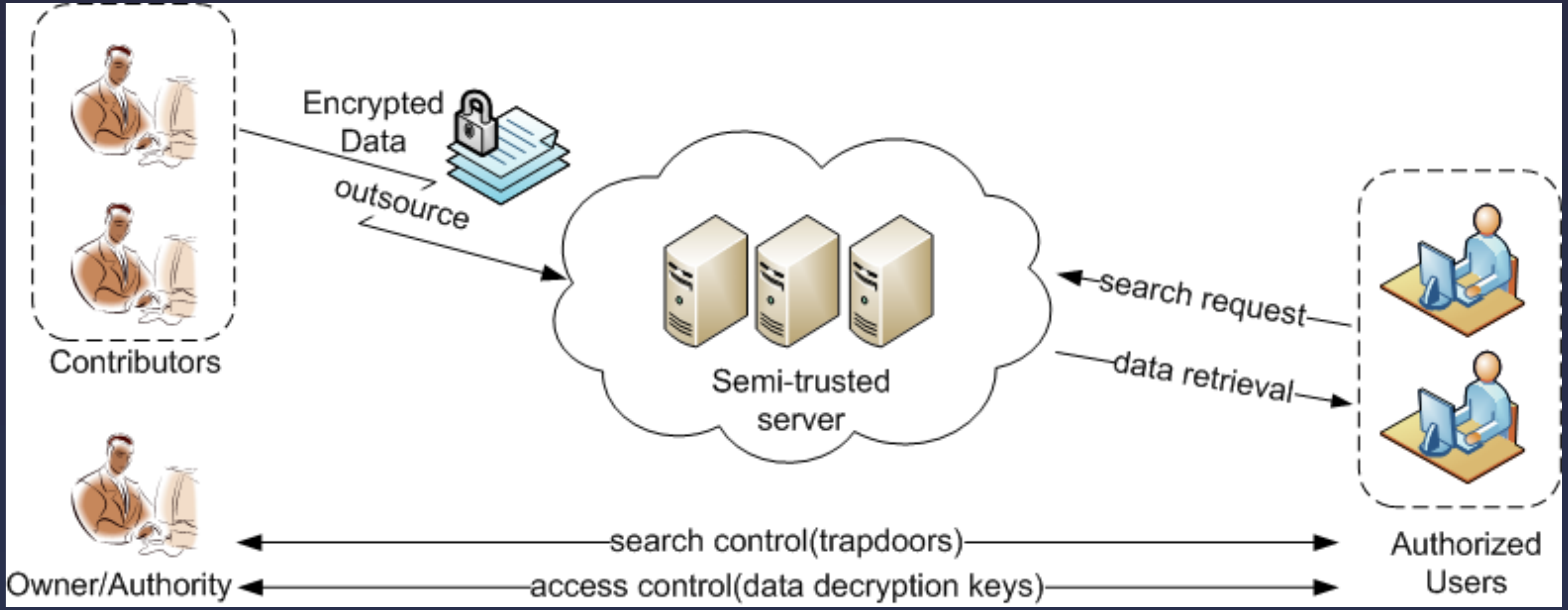


- Data contributor (data owner) encrypts and outsources data to semi-trusted server;
- Trusted authority (data owner) gives authorized users the search capability (e.g. trapdoors);
- Authorized users send search capability to server who will execute search over encrypted data;

Single Data Contributor

- Applications
  - Private email -- email server
  - Remote storage -- storage server
  - Medical records -- data server
  - Public health monitoring
  - Stock trading via semi-trusted brokers

Multiple Data Contributors

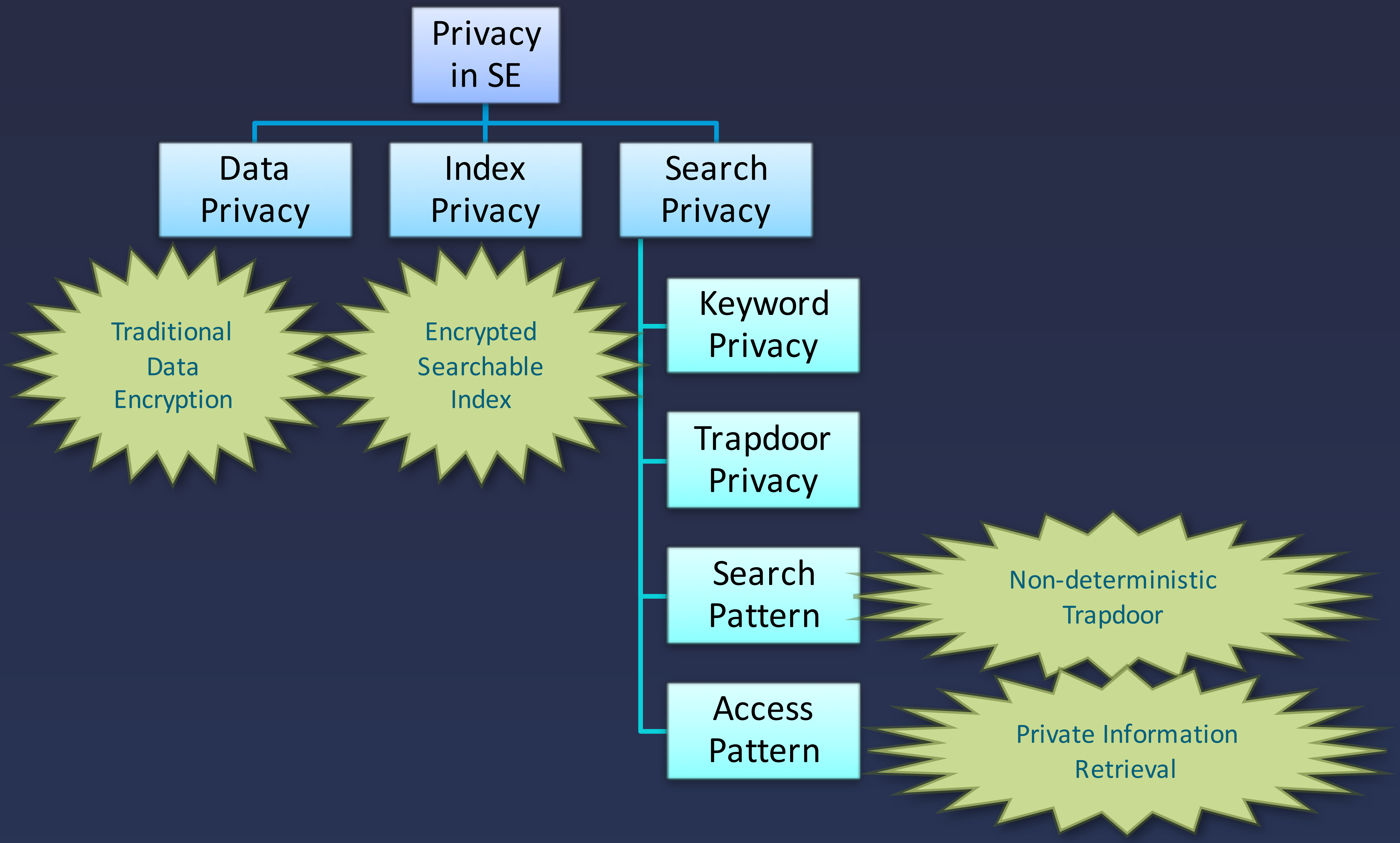


- Data contributors encrypts and outsources data to semi-trusted server;
- Trusted authority (data owner) gives authorized users the search capability (e.g. trapdoors);
- Authorized users send search capability to server who will execute search over encrypted data;

Multiple Data Contributors

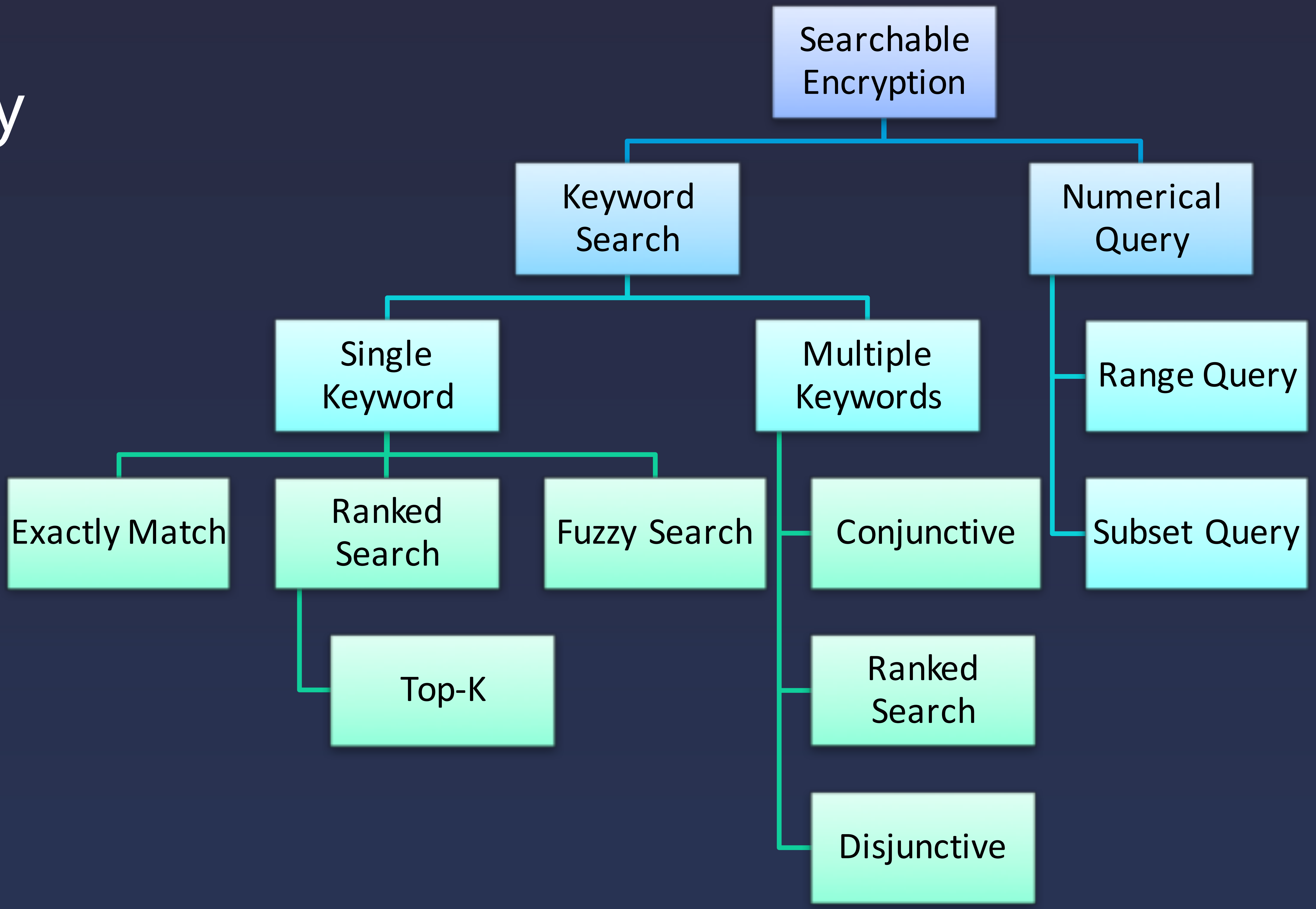
- Applications
  - Email server, Email gateway
  - Credit card payment gateway
  - Database
  - Medical records
  - Audit logs -- network, financial
  - network gateway/financial institutions, authorized auditor

Privacy Issues



Existing Work

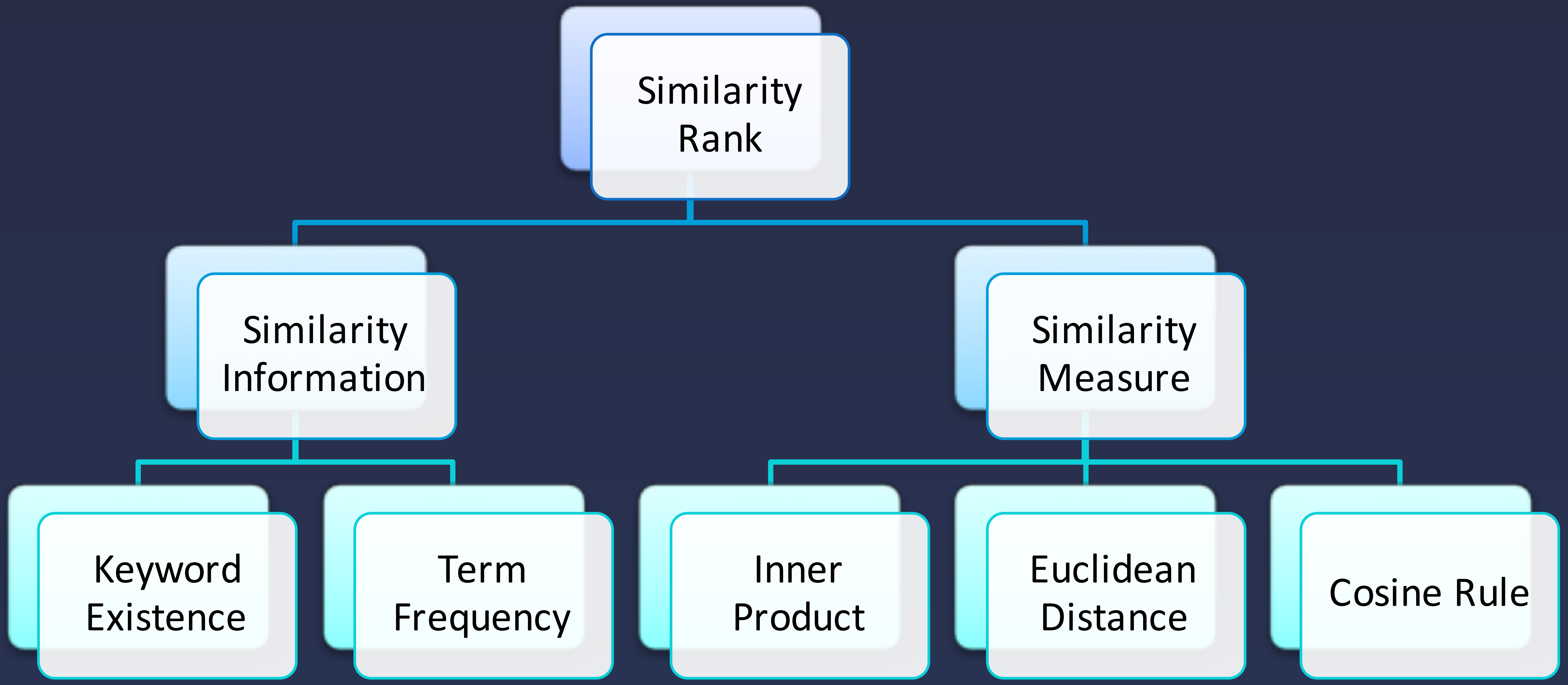
- Functionality



ArchSummit 全球架构师峰会 2017
主办方 Geekbang> 极客邦科技 InfoQ

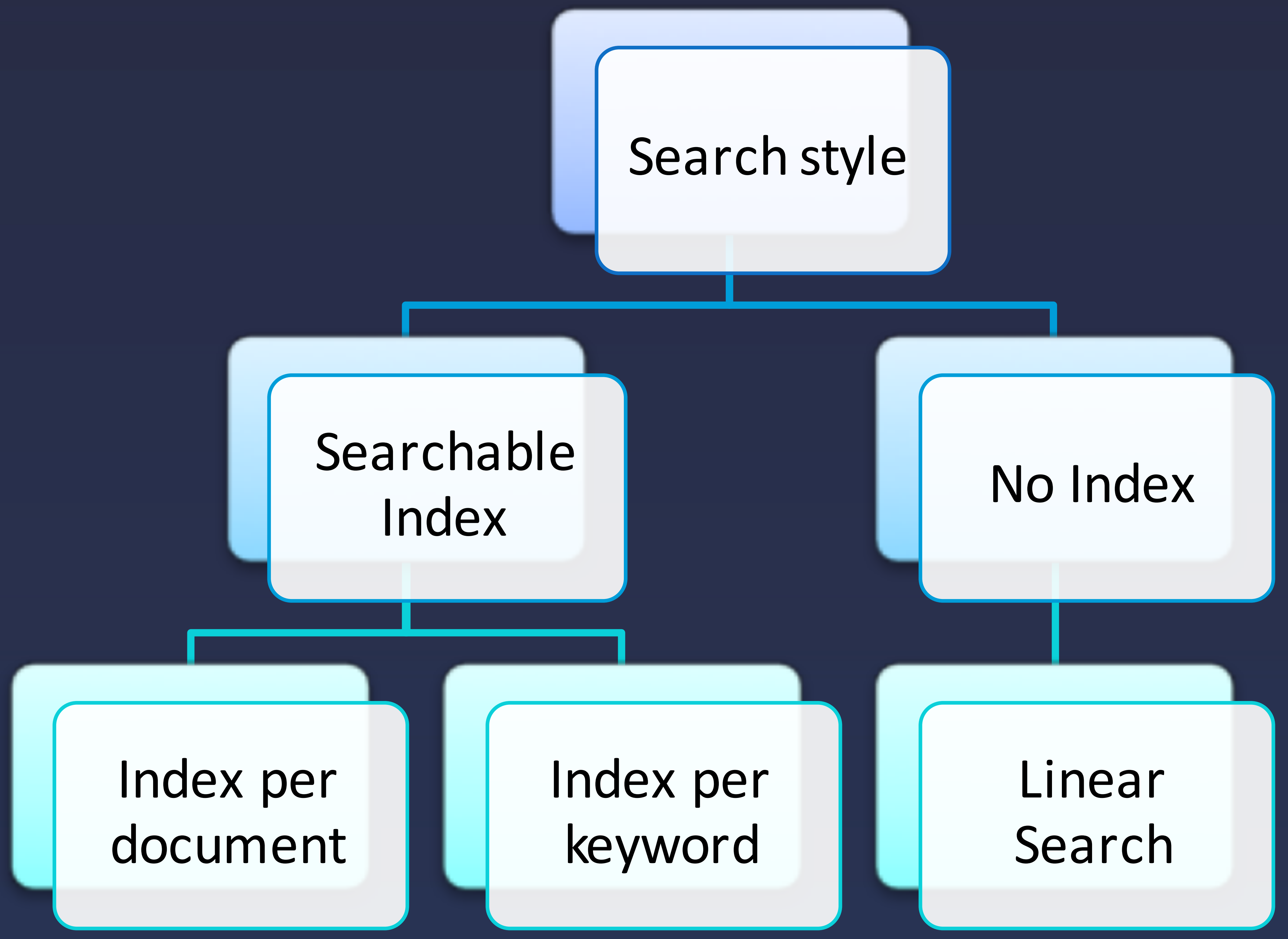
Existing Work

- IR Rank Technique



Existing Work

- Index construction



Existing Work

- Crypto Technique

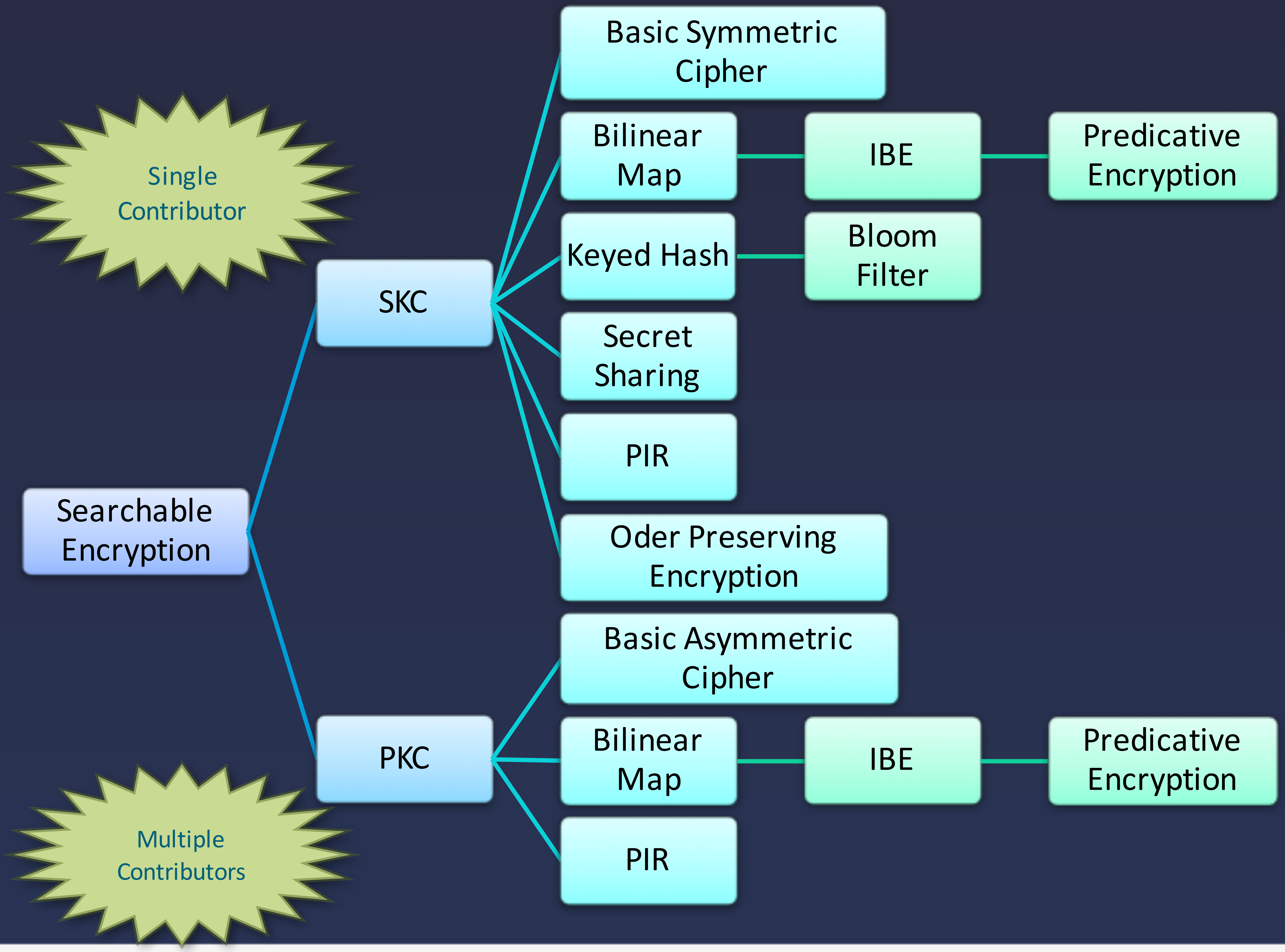


TABLE OF CONTENTS 大纲

Predicate Encryption

- Traditional encryptions: only owner of secret key can decrypt
- Attribute-based Encryption(ABE): fine-grained access control
  - E.g., Ciphertext-Policy based ABE
    - Access policy embedded in ciphertext
    - Key associated with attributes
    - Ciphertext could be decrypted if key’ s attributes satisfy access policy

Predicate Encryption

- Predicate Encryption:
  - plaintext m, attribute I -> ciphertext C
  - predicate/function $f_y()$ -> trapdoor/token/key $F_y()$
  - cipher text C could be decrypted as m iff $F_y(C) = f_y(I) = 1$

Predicate Encryption

- Predicate-only Encryption:
  - attribute I -> ciphertext C
  - predicate/function $f_y()$ -> trapdoor/token/key $F_y()$
  - $F_y(C)$ =1 iff $f_y(I) = 1$

Predicate Encryption

- Existing works:
  - Identity-based encryption: equality tests
  - Attribute-based encryption: conjunctions, range queries
  - Predicate encryption: disjunctions, inner products, etc
- Cons: computation complexity
  - bilinear map: $e:G\times G\rightarrow G_T$, $e(u^a, v^b) = e(u, v)^{ab}$

THANKS

让创新技术推动社会进步

HELP TO BUILD A BETTER SOCIETY WITH INNOVATIVE TECHNOLOGIES

Geekbang.

极客邦科技

InfoQ

专注中高端技术人员的技术媒体

EGO NETWORKS EXTRA GEEKS' ORGANIZATION

高端技术人员学习型社交平台

StuQ 斯达克学院

实践驱动的 IT 教育平台

地址：北京市朝阳区洛娃大厦C座8层1801室　　网址：www.geekbang.org