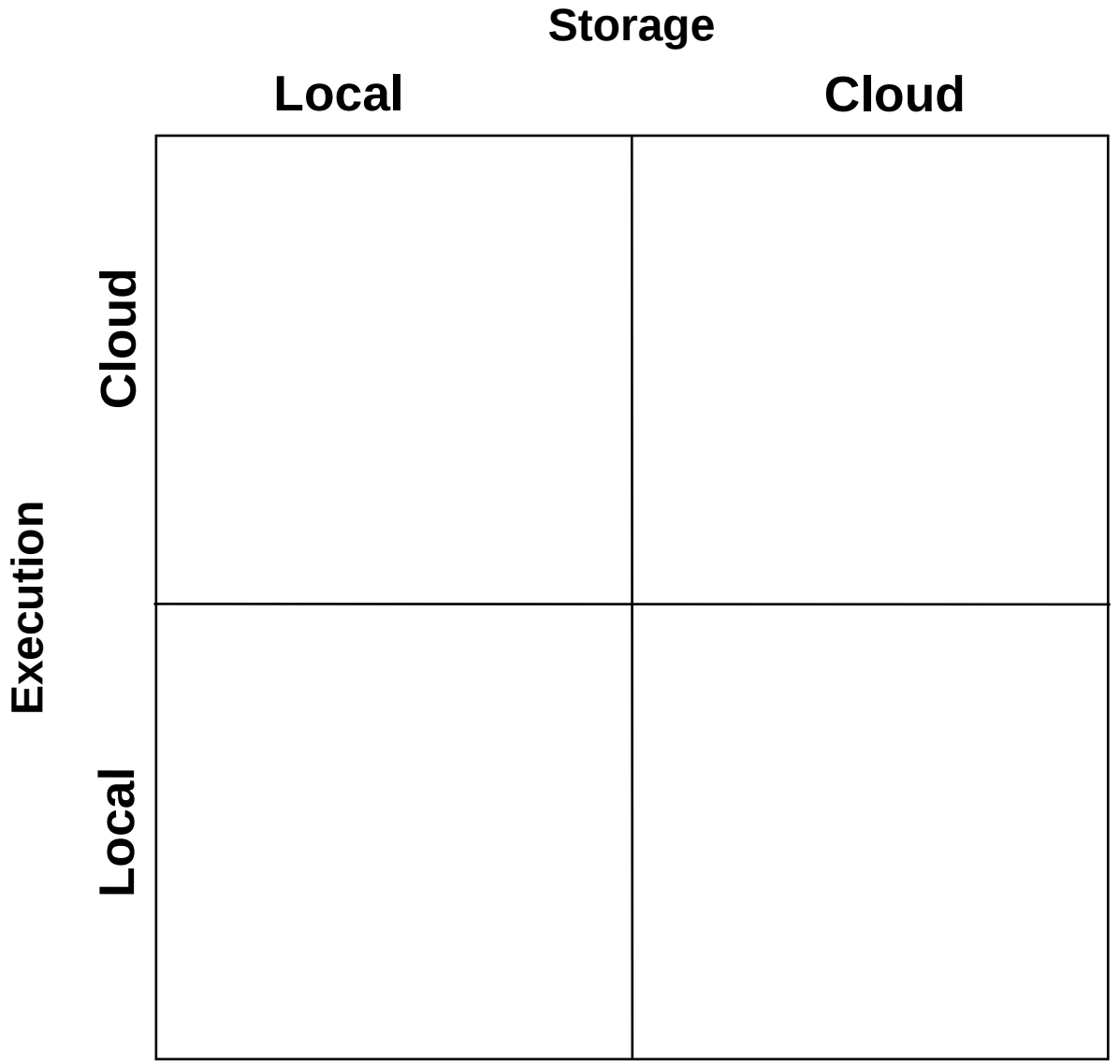
The Case for Content Search of VM Clouds

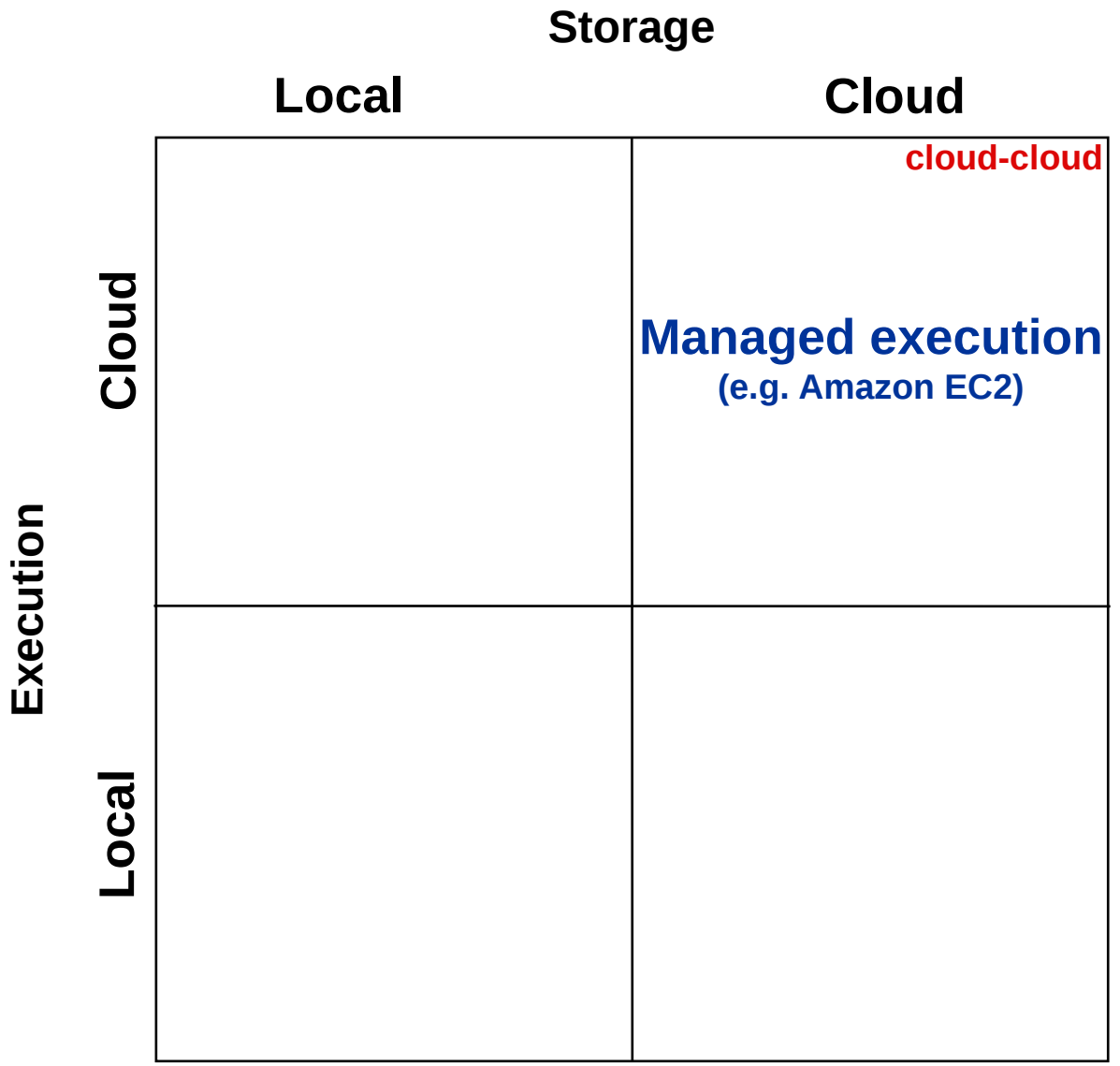
Mahadev Satyanarayan, Wolfgang Richter, Glenn Ammons[†], Jan Harkes and Adam Goode

Carnegie Mellon University and [†]IBM Research

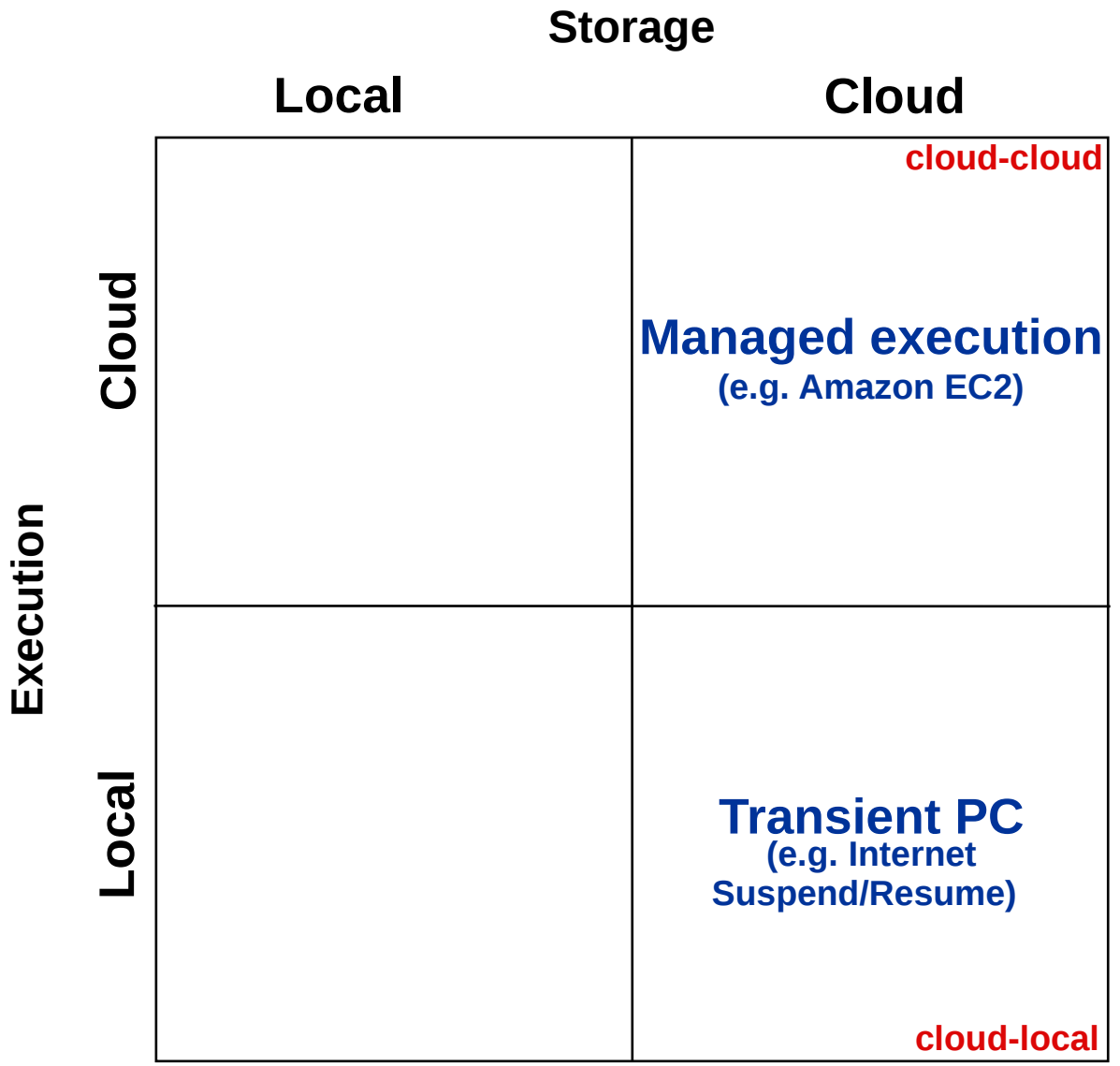
Types of VM/Cloud Computing



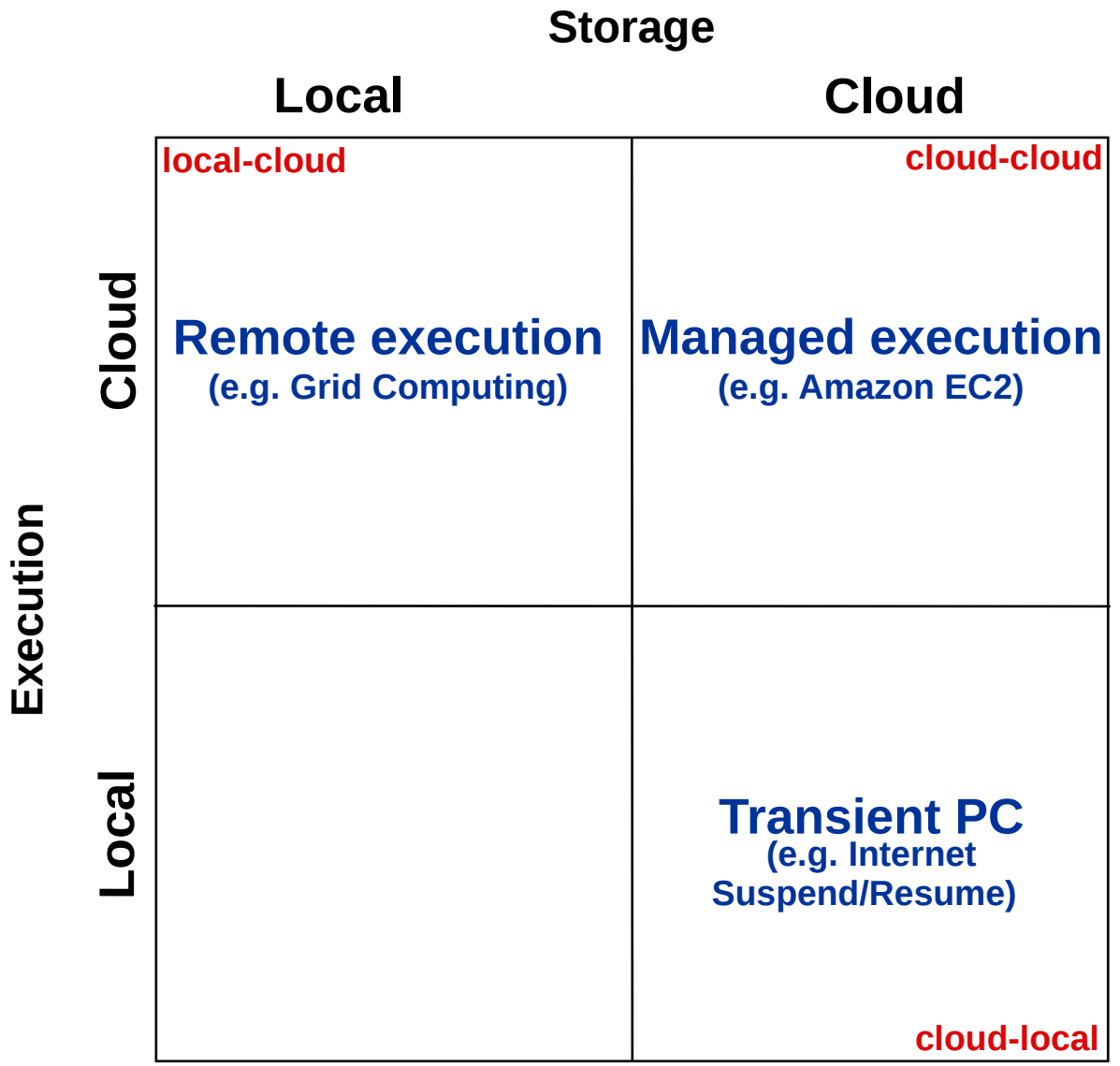
Types of VM/Cloud Computing



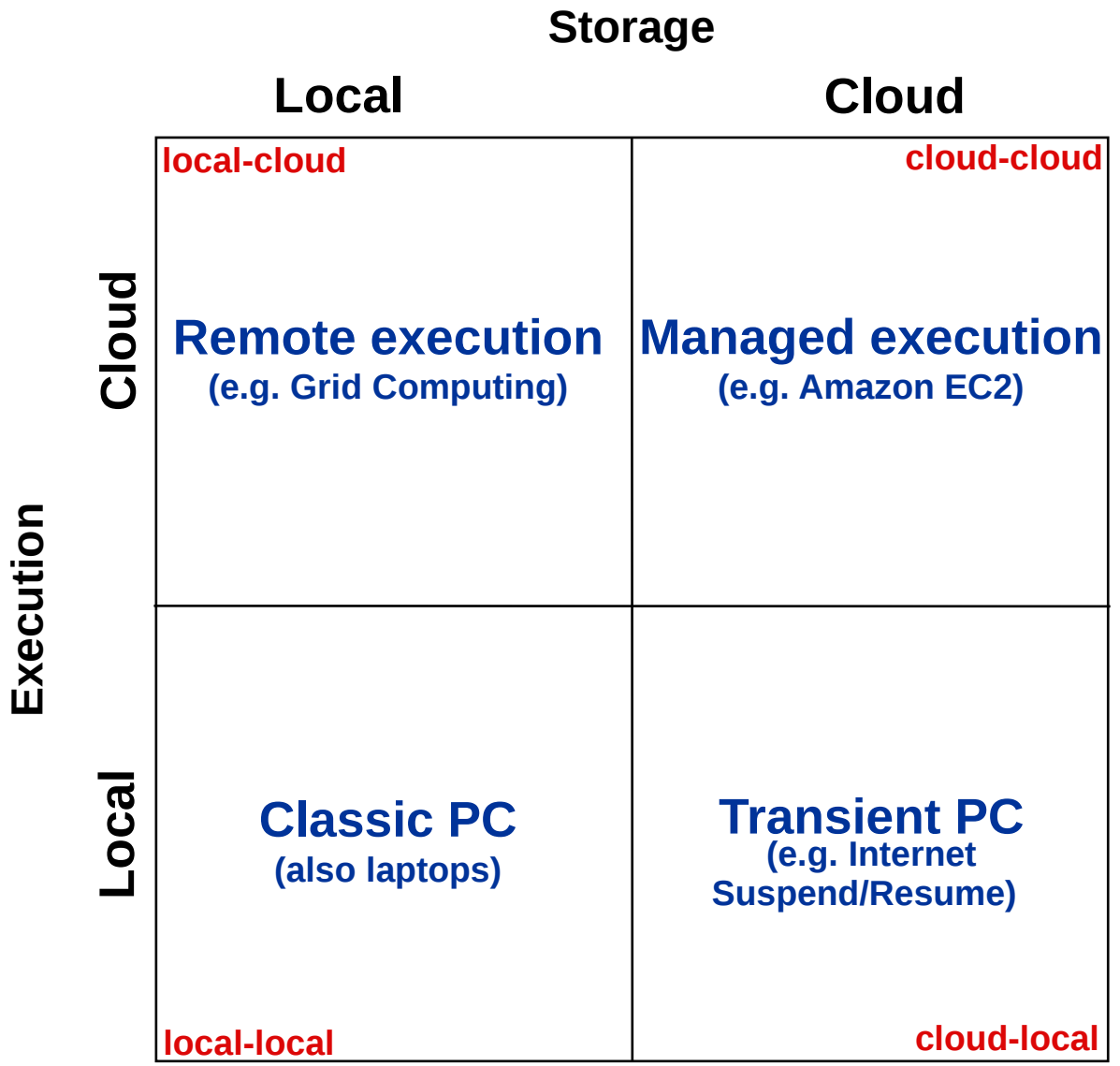
Types of VM/Cloud Computing



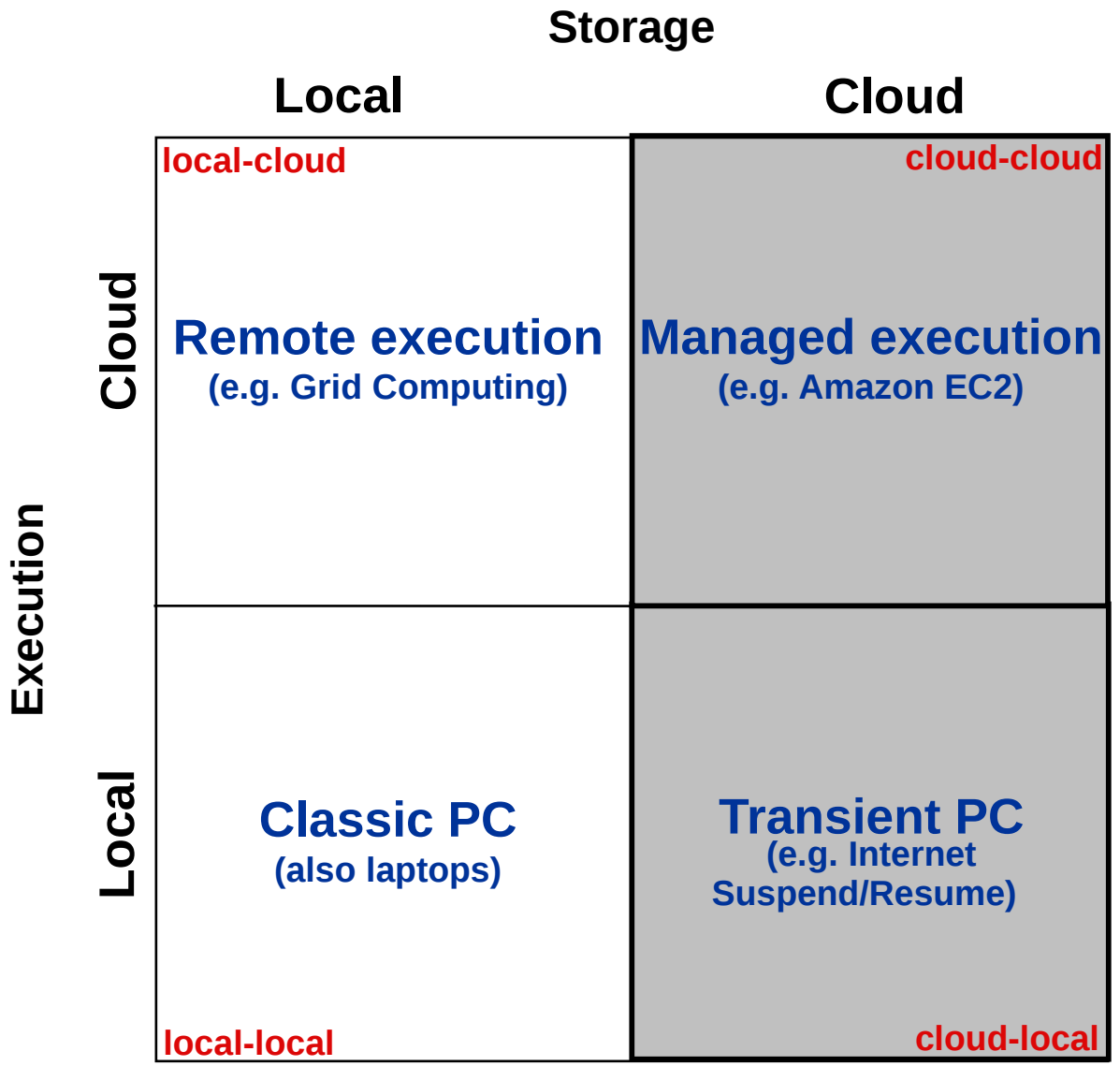
Types of VM/Cloud Computing



Types of VM/Cloud Computing



Types of VM/Cloud Computing



Content **is** Important

- Specific versions of software
  - Tool chains, DLL's
- Proprietary Data
- Custom User Data
- Research Experiment Setups

Potential Applications

- Graphic Design – **Copyright Infringement**
- Corporate Policy – **Software Updates**, **Licenses**
- Software Development – **Debugging**

Search **Requirements**

- **Content-Based** Searching
  - Not meta-data alone
- **Domain-Dependent** Queries
  - Sensitive to type of data and search primitives
- **Iterative** Search Workflow
  - More than data mining

Discard-Based Search: Diamond

User

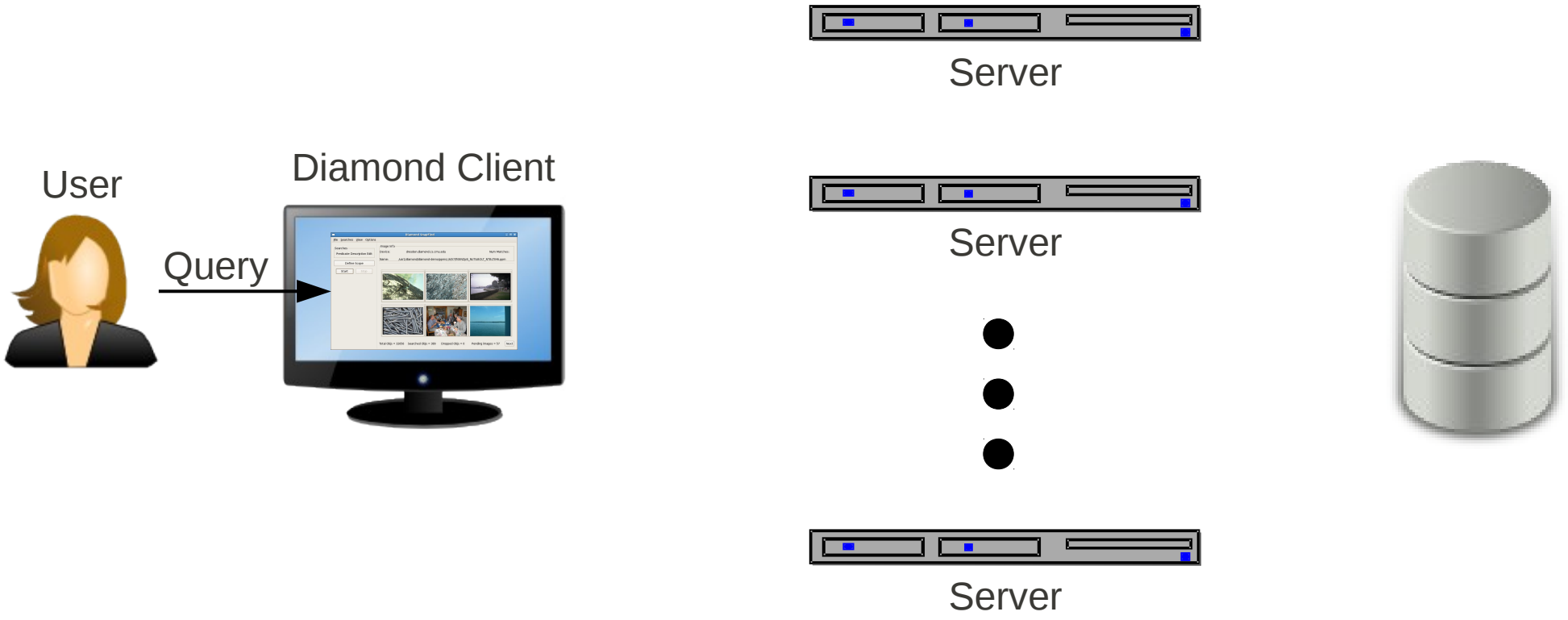
Diamond Client

Server

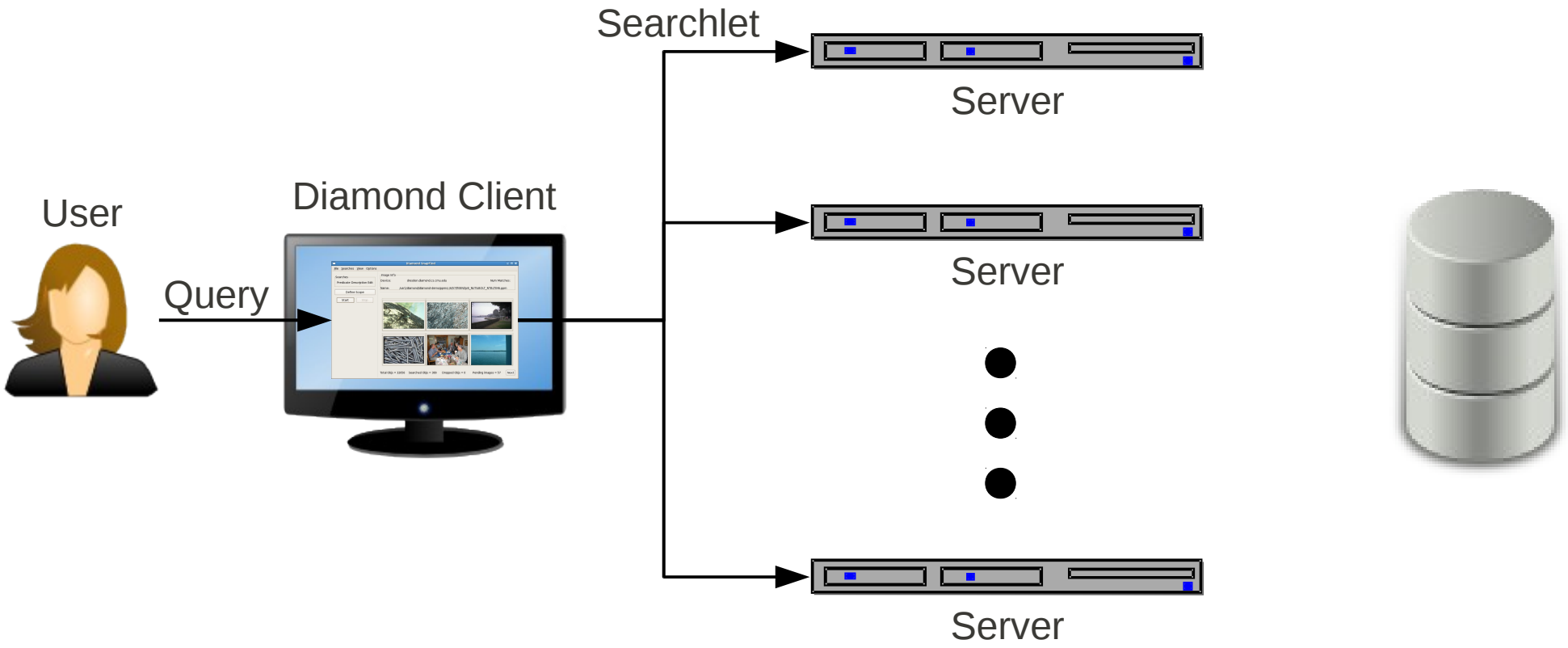
Server

Server

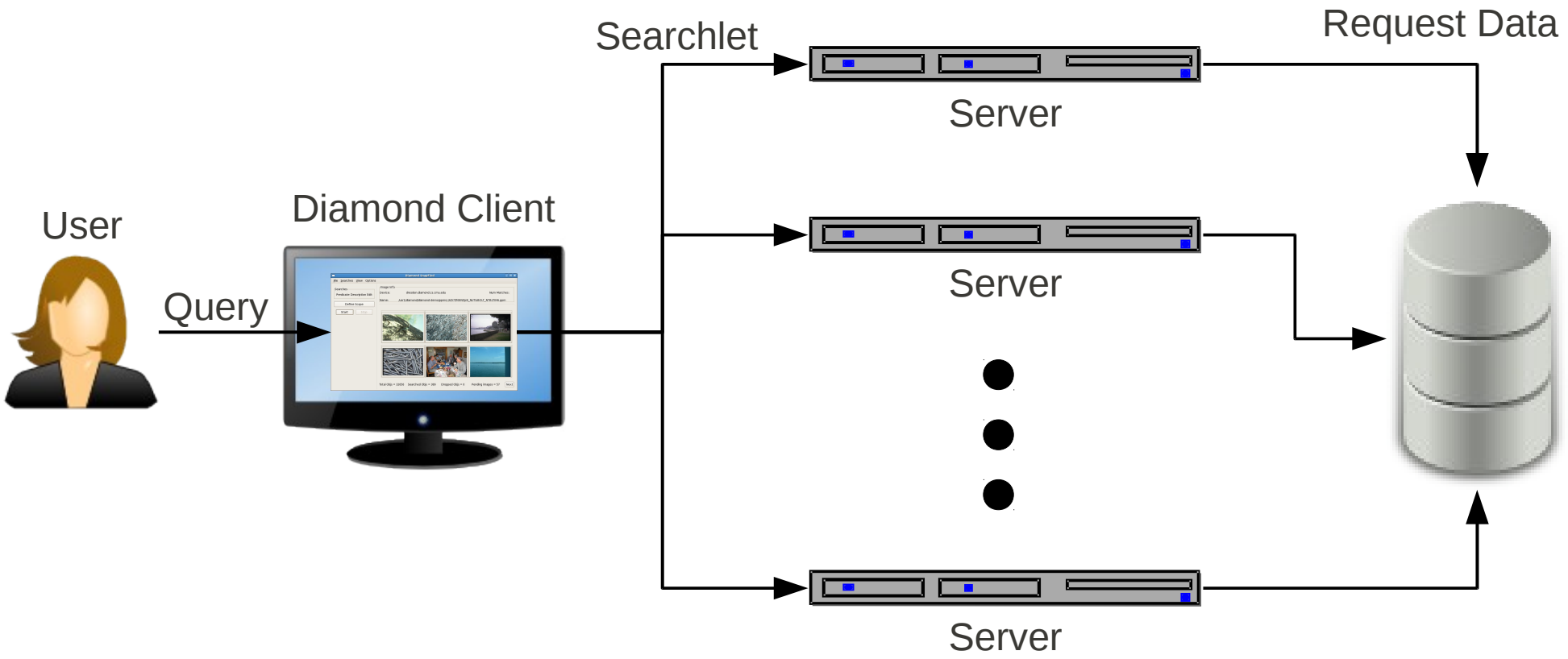
Discard-Based Search: Diamond



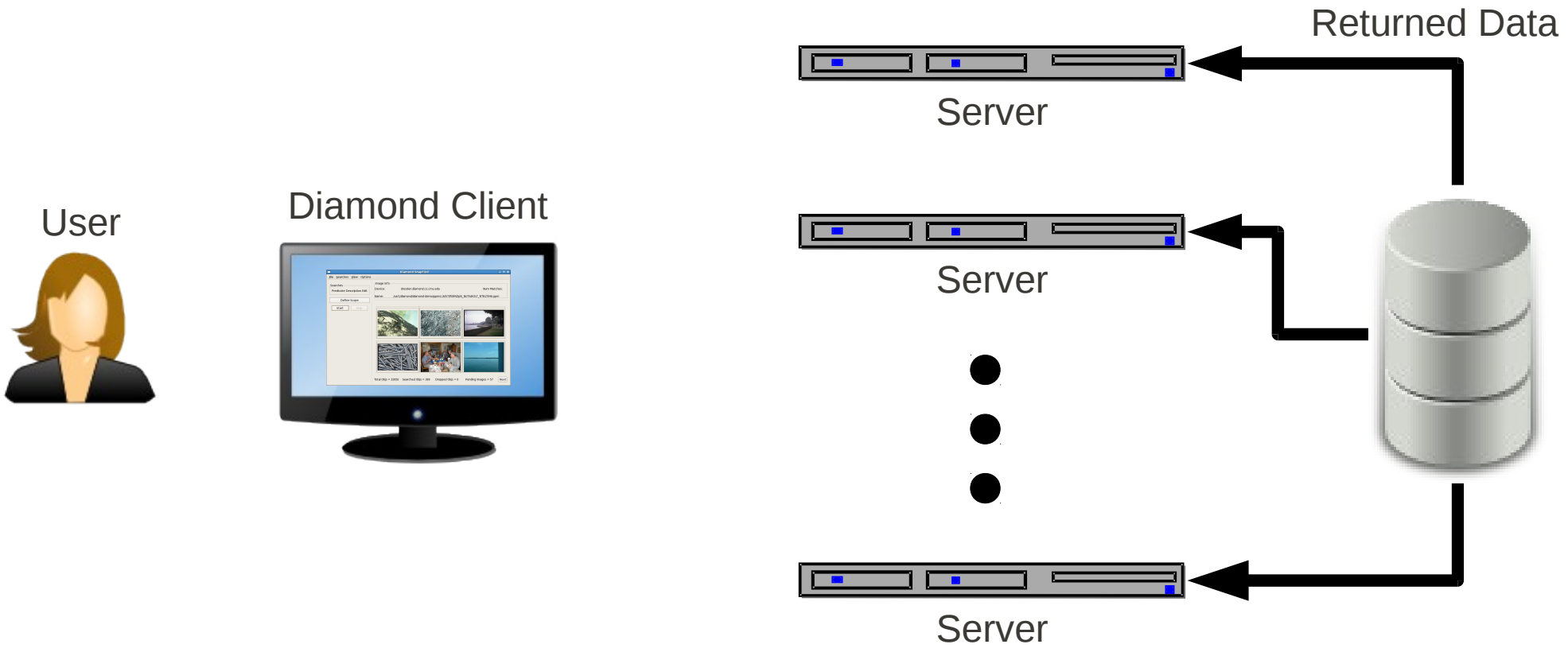
Discard-Based Search: Diamond



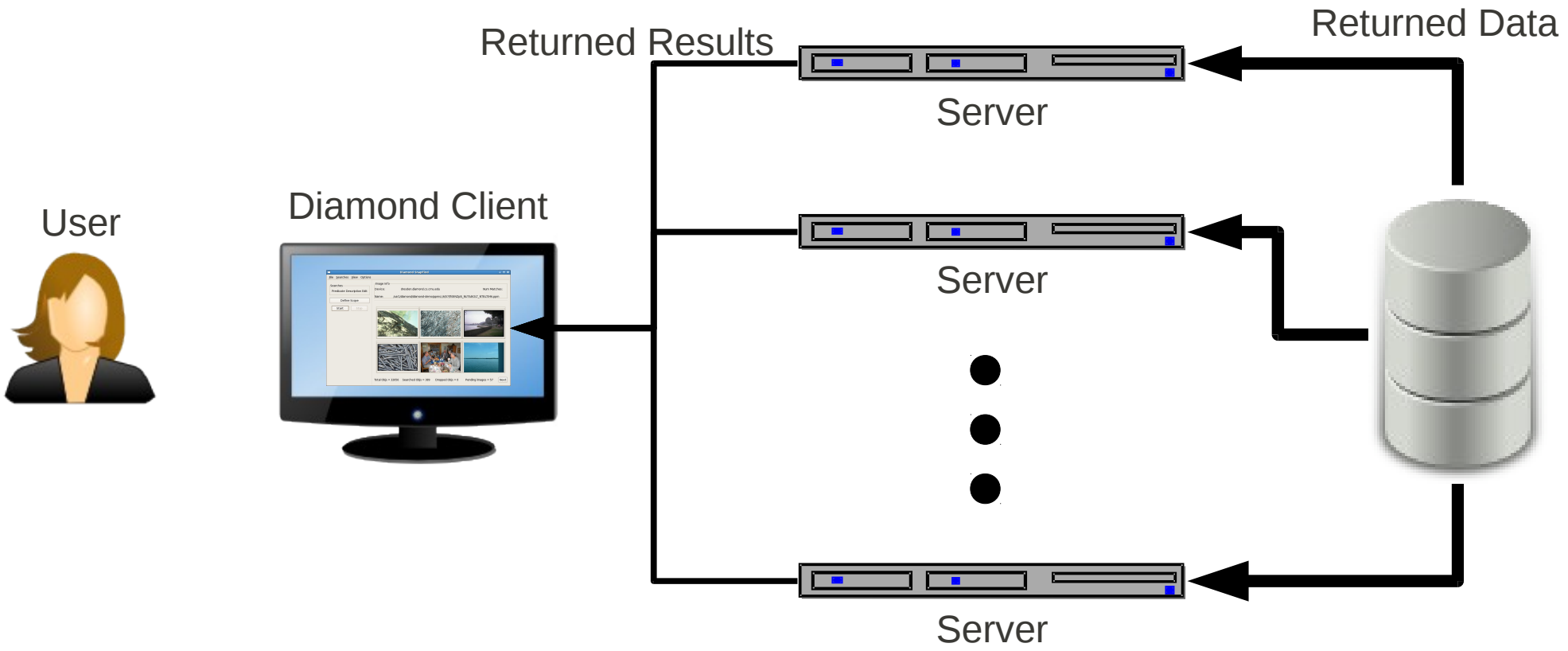
Discard-Based Search: Diamond



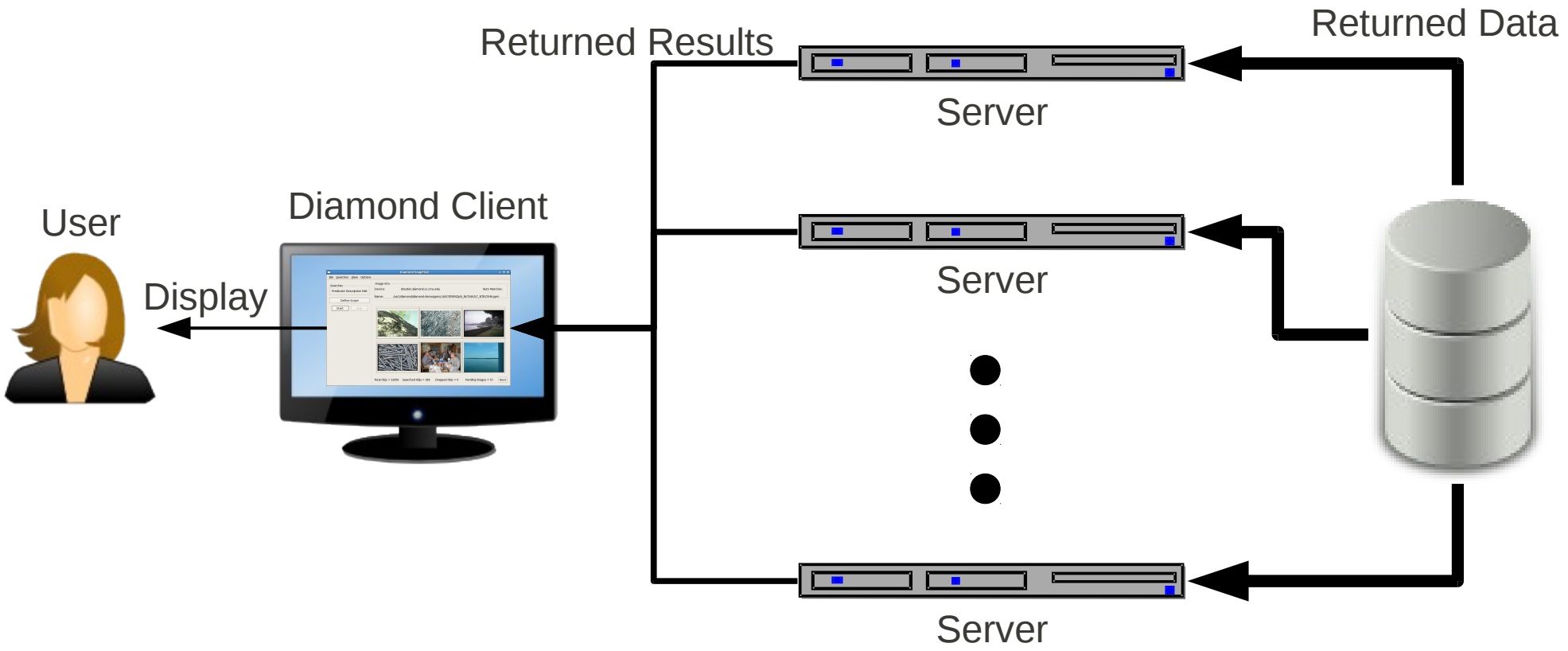
Discard-Based Search: Diamond



Discard-Based Search: Diamond



Discard-Based Search: Diamond



Filters and **Searchlets**

- **Filter** – domain-specific executable code
  - Inherent parallelism
  - **Temporal Locality** enables **just-in-time indexing**
- **Searchlet** – Collection of parameterized filters
- Returned **Result** – Pass all filters in a searchlet

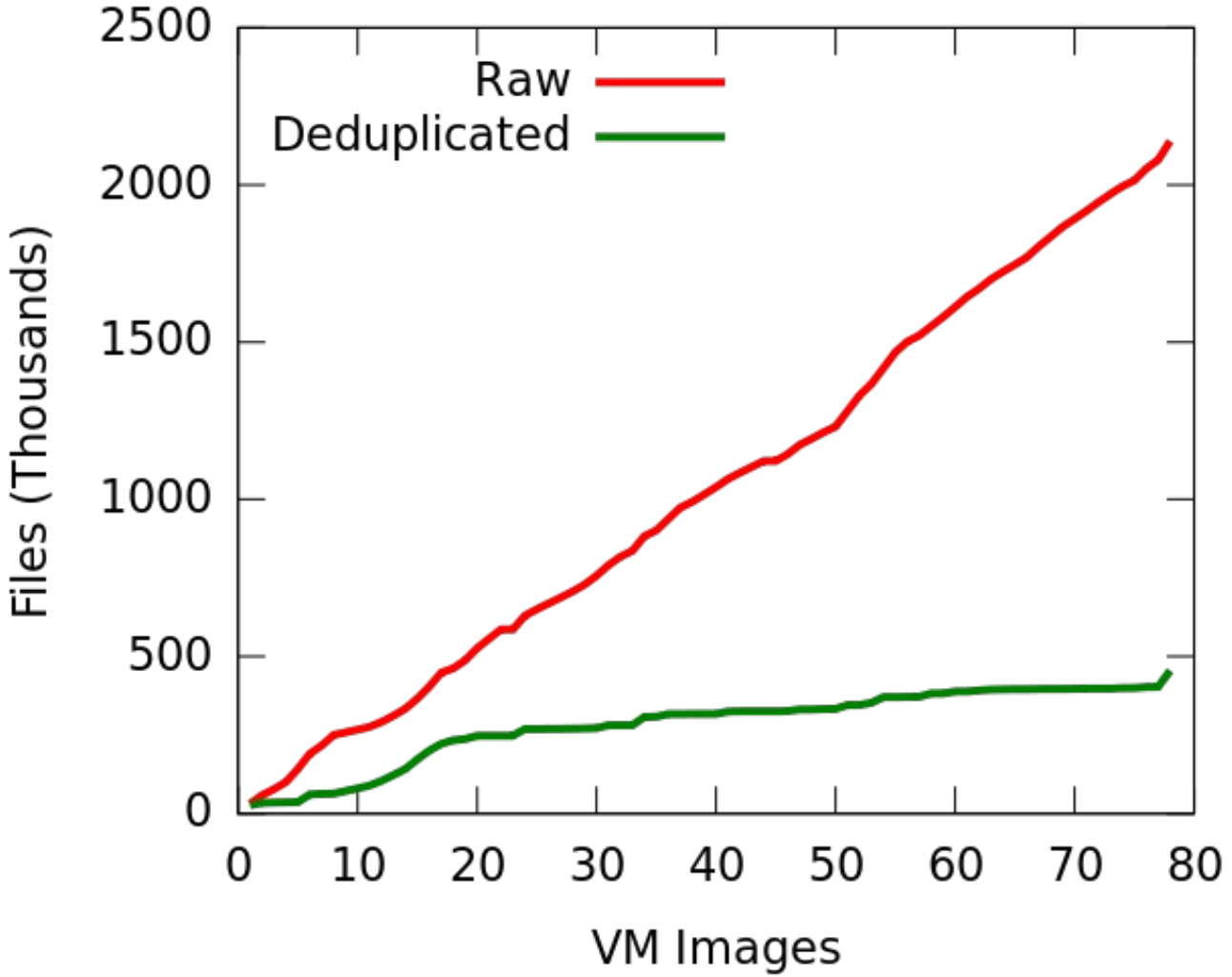
VM Growth

- Worldwide Pervasive VM Technology
- EC2, RC2, Eucalyptus, ISR, MokaFive...
- 1,000 Employees, 1,000 Daily Snapshots

VM Search Solved?

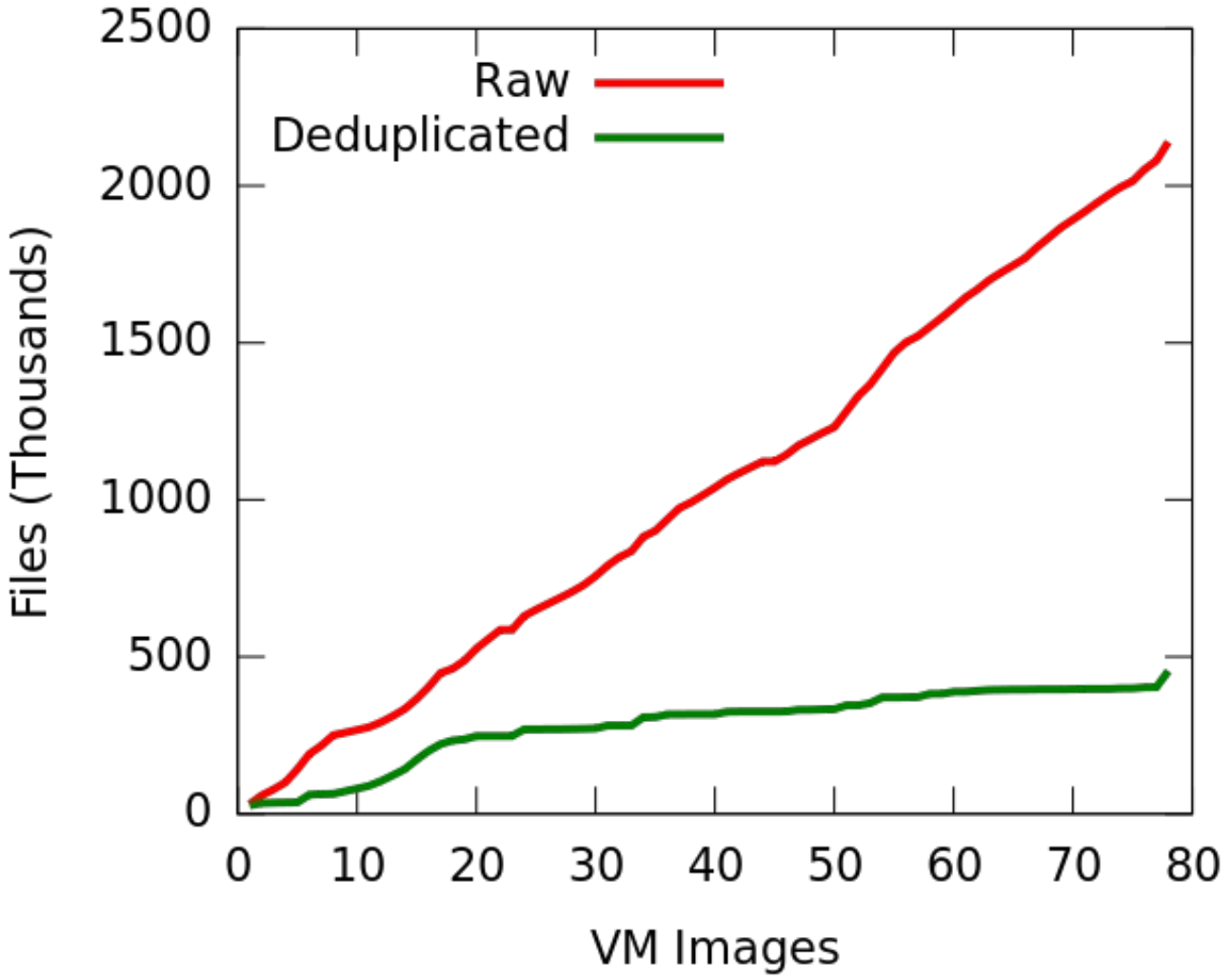
- VM's generally multi-gigabyte in size
- 1000's of VM's Implies Terabytes of Data
- **Key Insight**: VM's often contain **duplicate data**
  - Same OS, Same Software, Same Data
  - Search the same file **once**, not **thousands** of times

Importance of **Deduplication**: Files



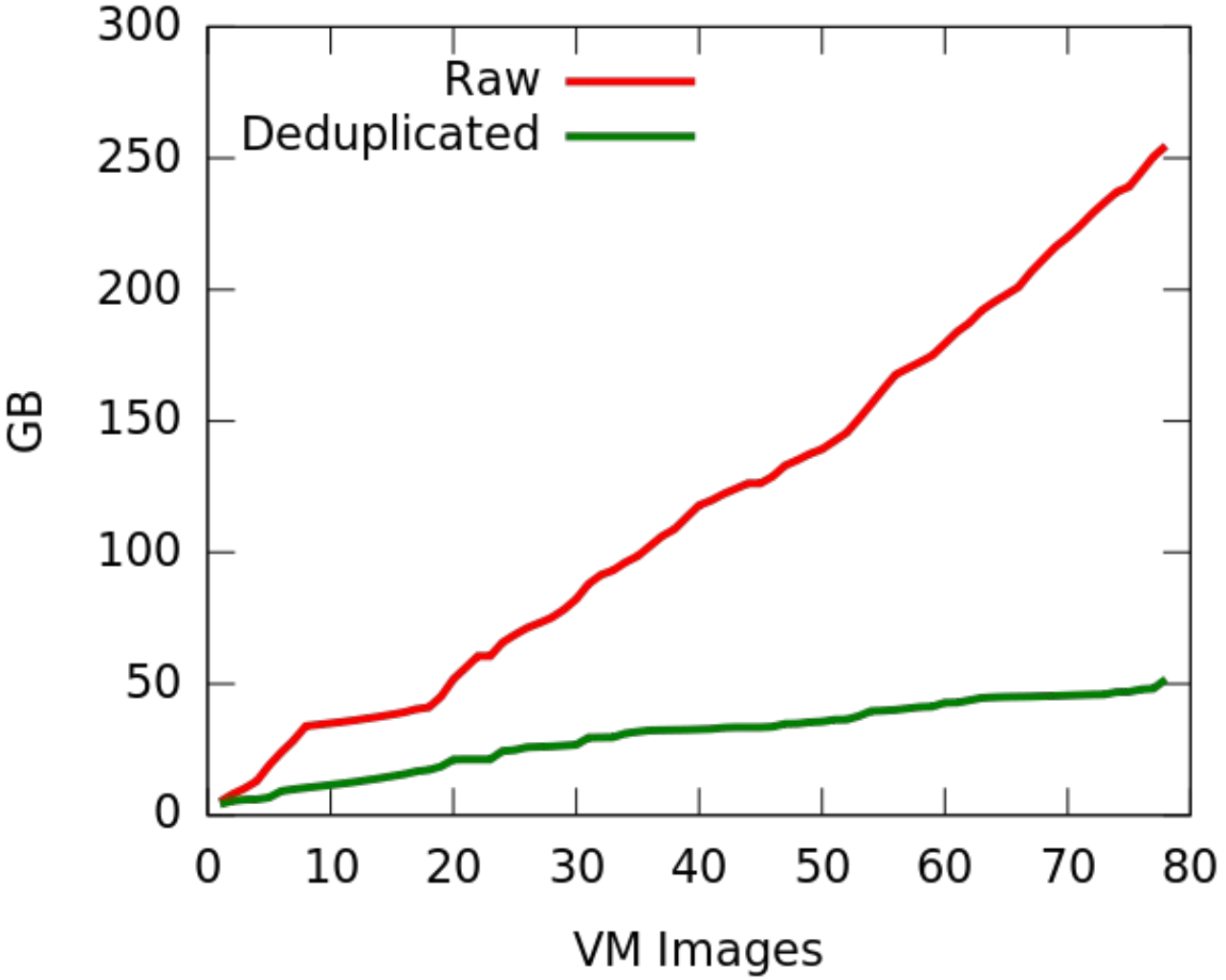
Data from 78 NCSU VCL parcels based on Windows XP

Importance of **Deduplication**: Files



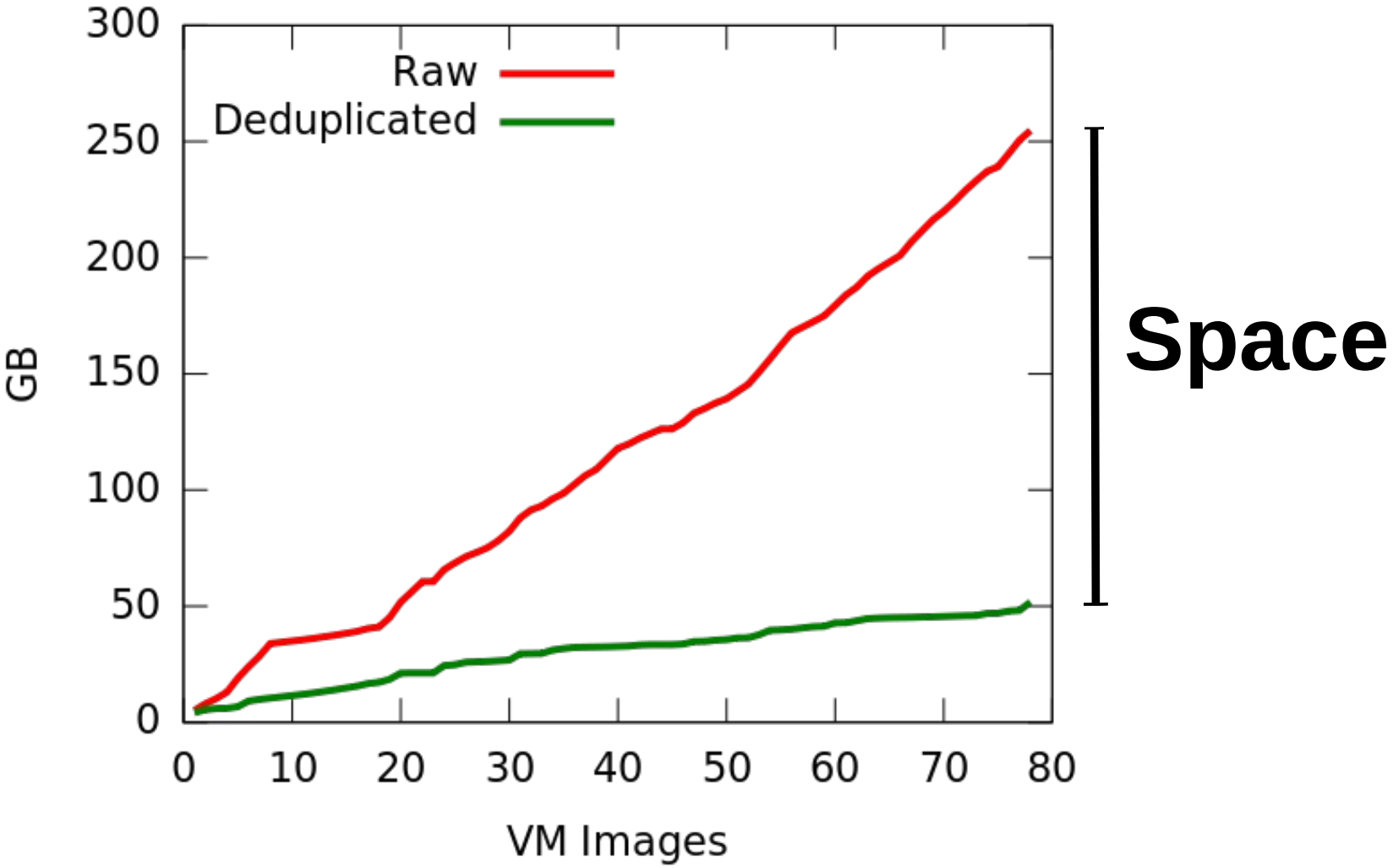
Data from 78 NCSU VCL parcels based on Windows XP

Importance of **Deduplication**: Bytes



Data from 78 NCSU VCL parcels based on Windows XP

Importance of **Deduplication**: Bytes



Data from 78 NCSU VCL parcels based on Windows XP

How to **Deduplicate** VM's?

- Virtual Machine Format Independent
- OS Independent
- File System Independent

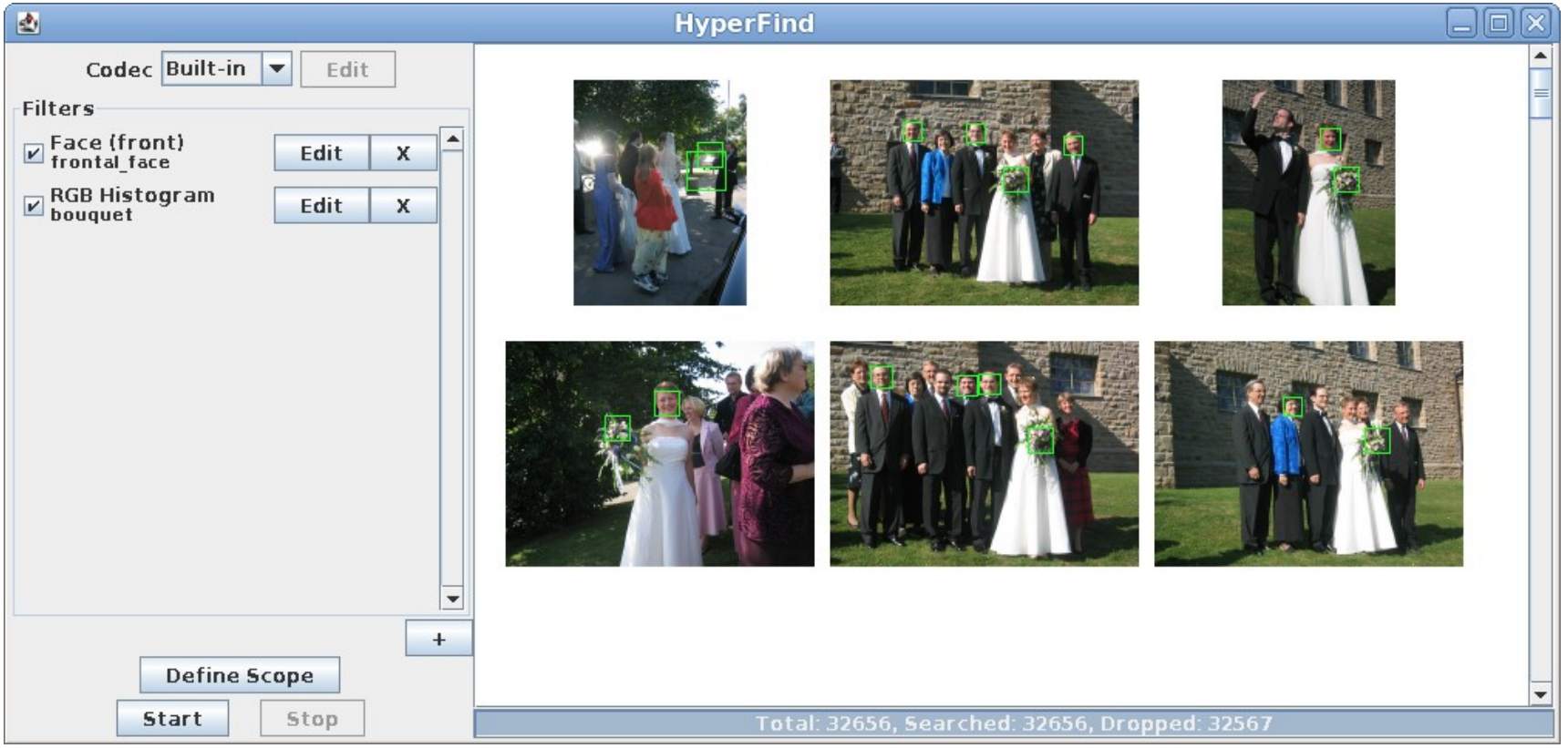
Deduplication Solution: **Mirage**

- IBM Research Project
- Parses Virtual Disk Partitions
- Handles Multiple File Systems
  - Currently Tested: **ext2**, **ext3**, and **NTFS**
- Extracts and Deduplicates Files

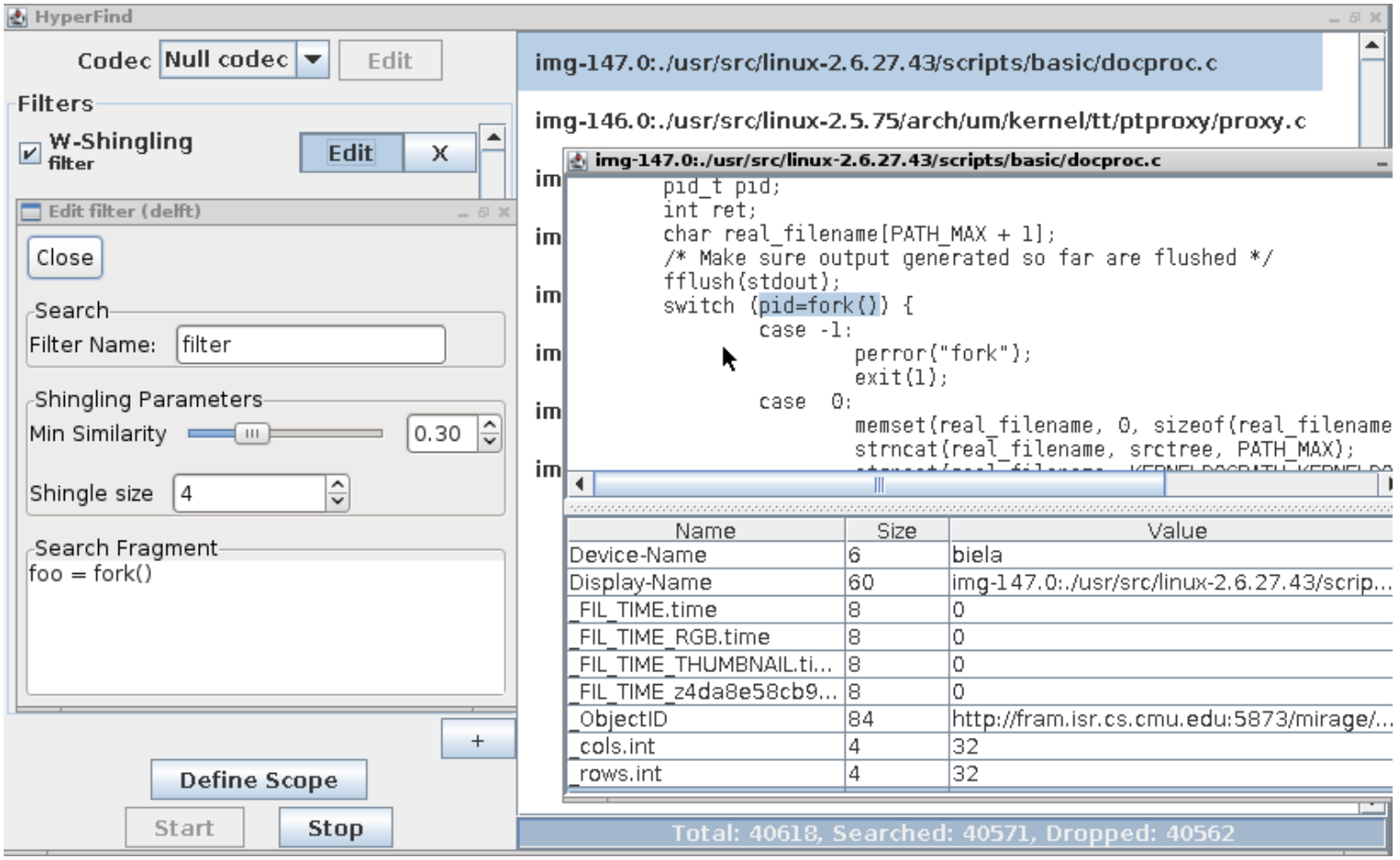
Integrating **Diamond+Mirage**

- Data Source Abstraction
- Database with File-Level Semantics
- Scoping Mechanism
  - Limit to certain Files
  - Provides Access Control

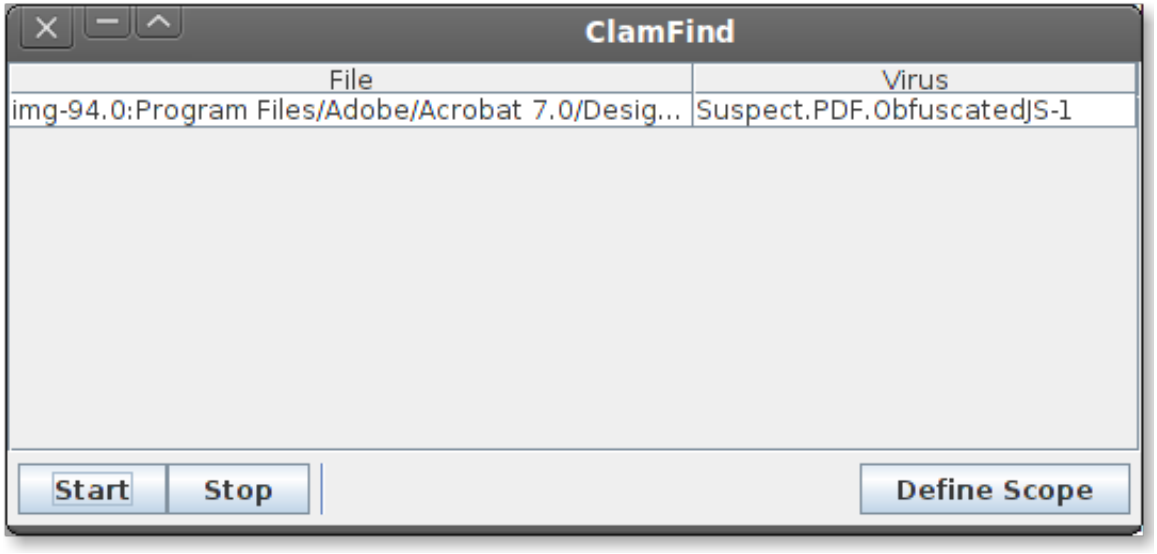
Application: **Image Search**



Application: **Source Code Search**



Application: **Vulnerability Search**



Conclusion

- Content search of VM's has many applications
- Number of VM's growing
- Proposed Solution: Diamond+Mirage
  - Addresses unique search requirements
  - Addresses deduplication

Questions?

Privacy? Ethics? Legality?