

The Case for Content Search of VM Clouds

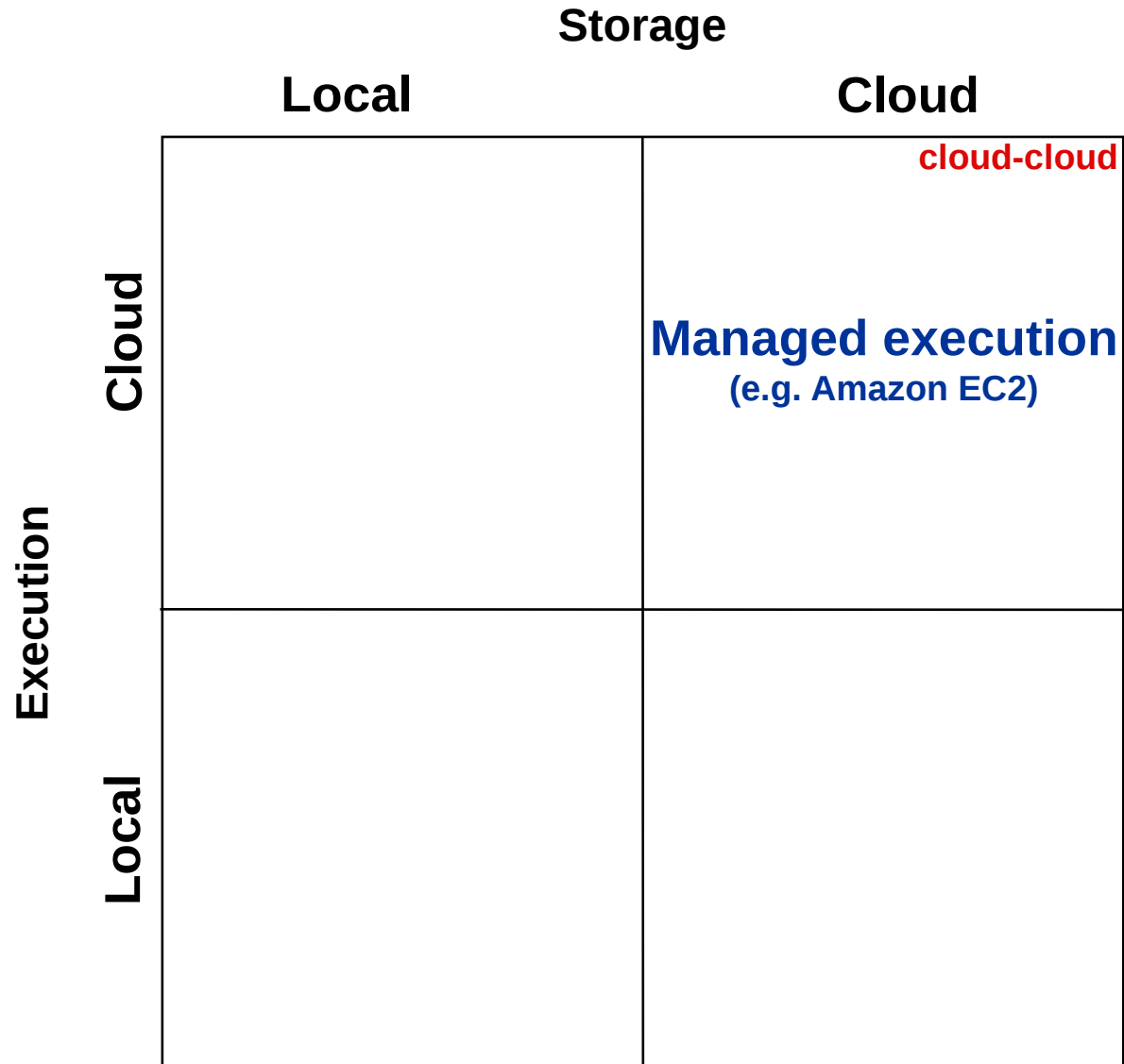
Mahadev Satyanarayan, Wolfgang Richter, Glenn Ammons[†], Jan Harkes and Adam Goode

Carnegie Mellon University and [†]IBM Research

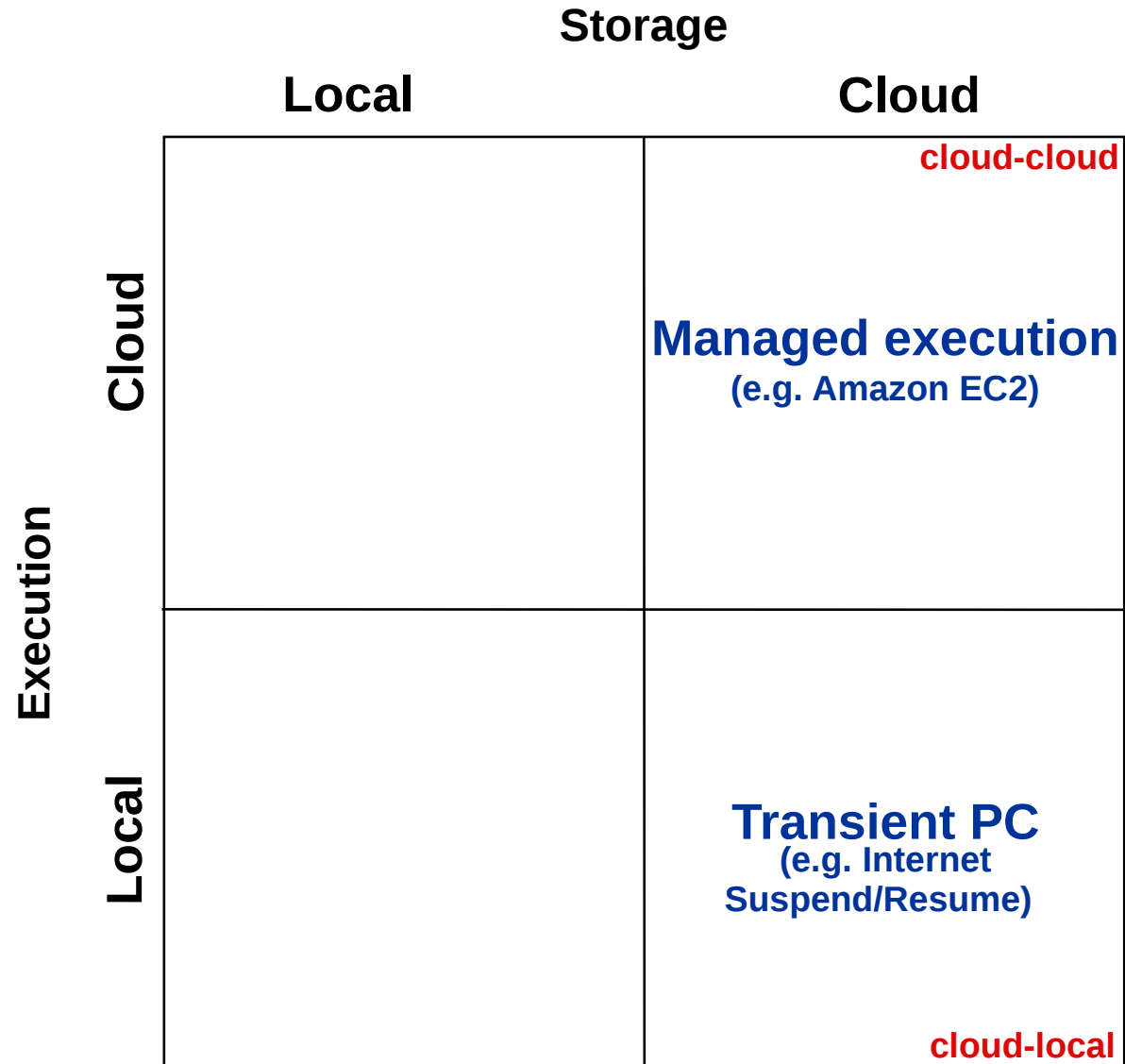
Types of VM/Cloud Computing

		Storage	
		Local	Cloud
Execution	Cloud		
	Local		

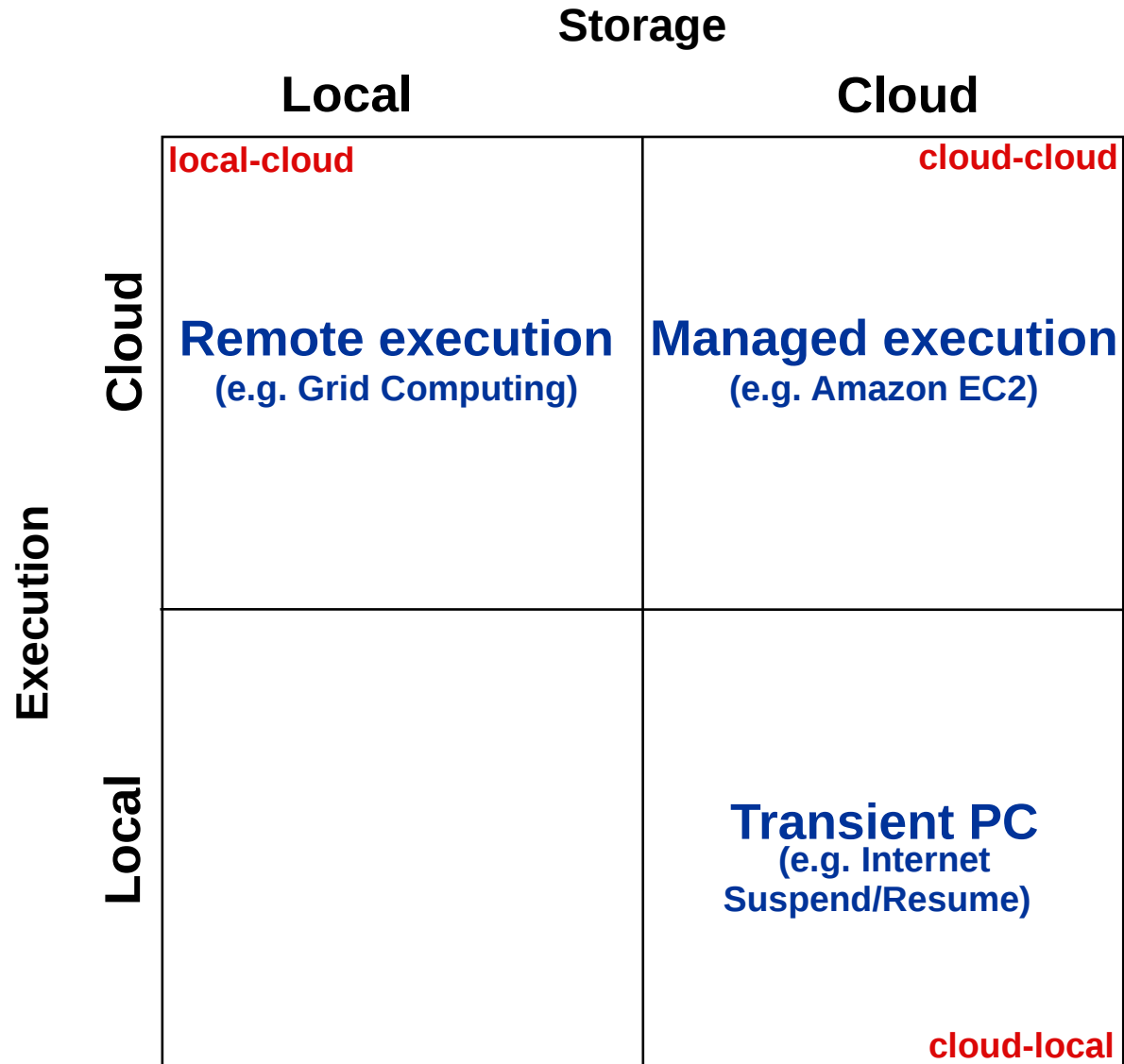
Types of VM/Cloud Computing



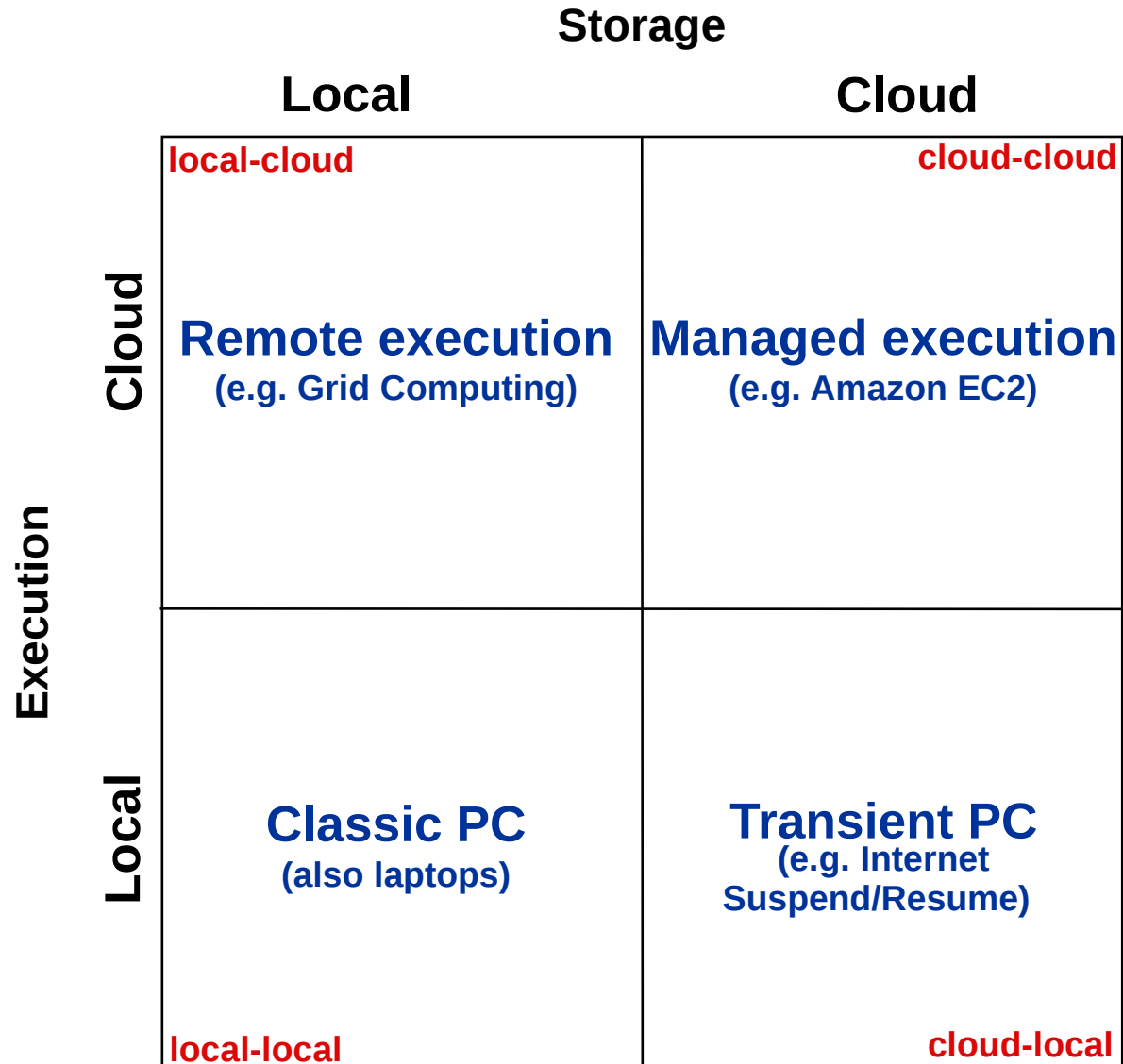
Types of VM/Cloud Computing



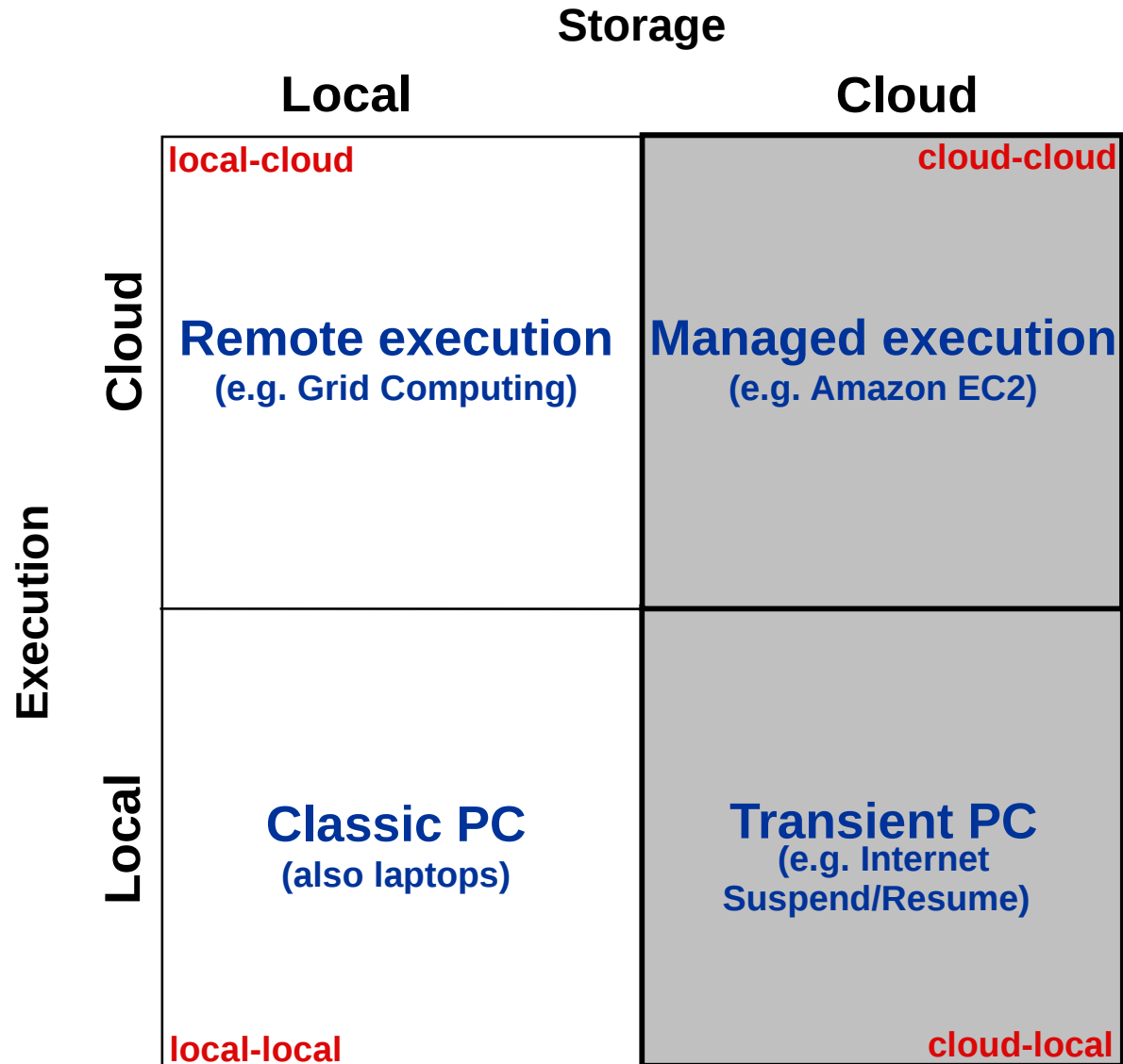
Types of VM/Cloud Computing



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**

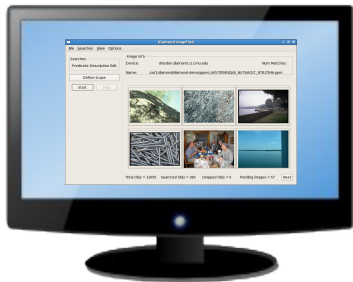
Search Requirements

- **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



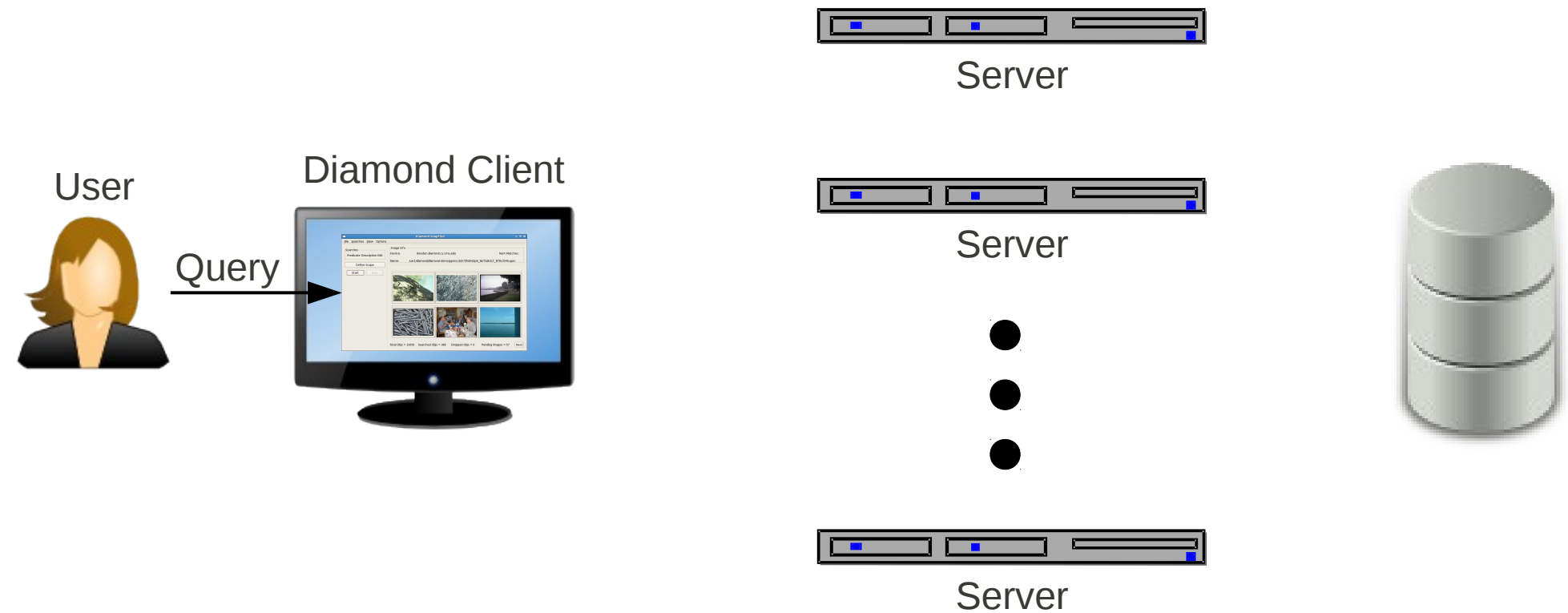
Server



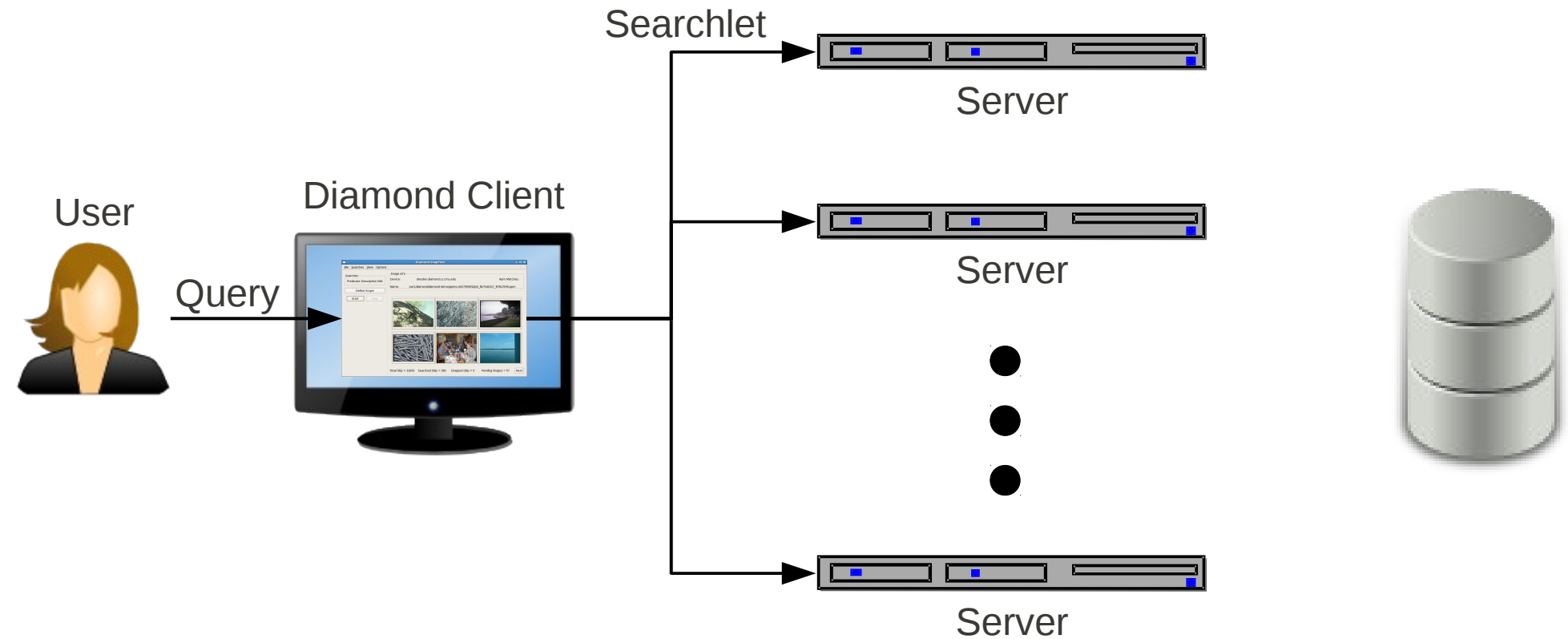
Server



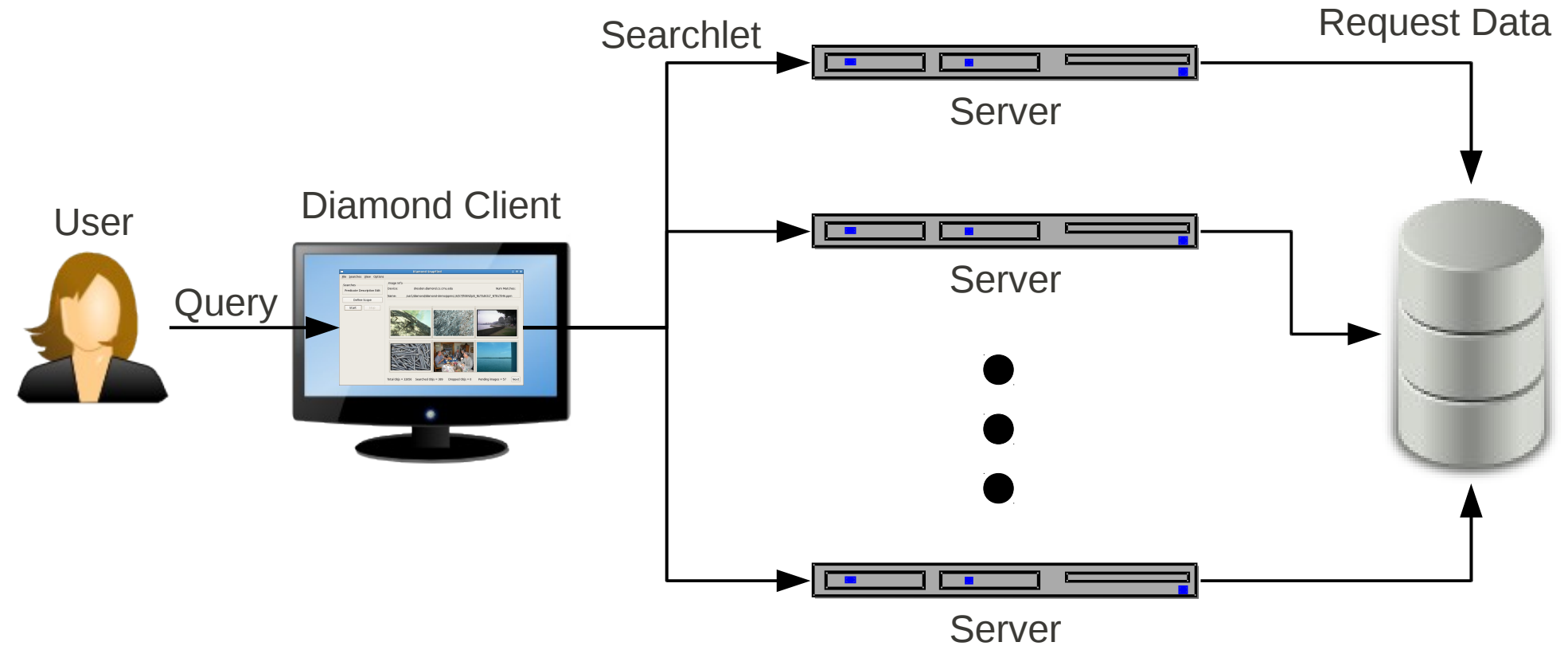
Discard-Based Search: Diamond



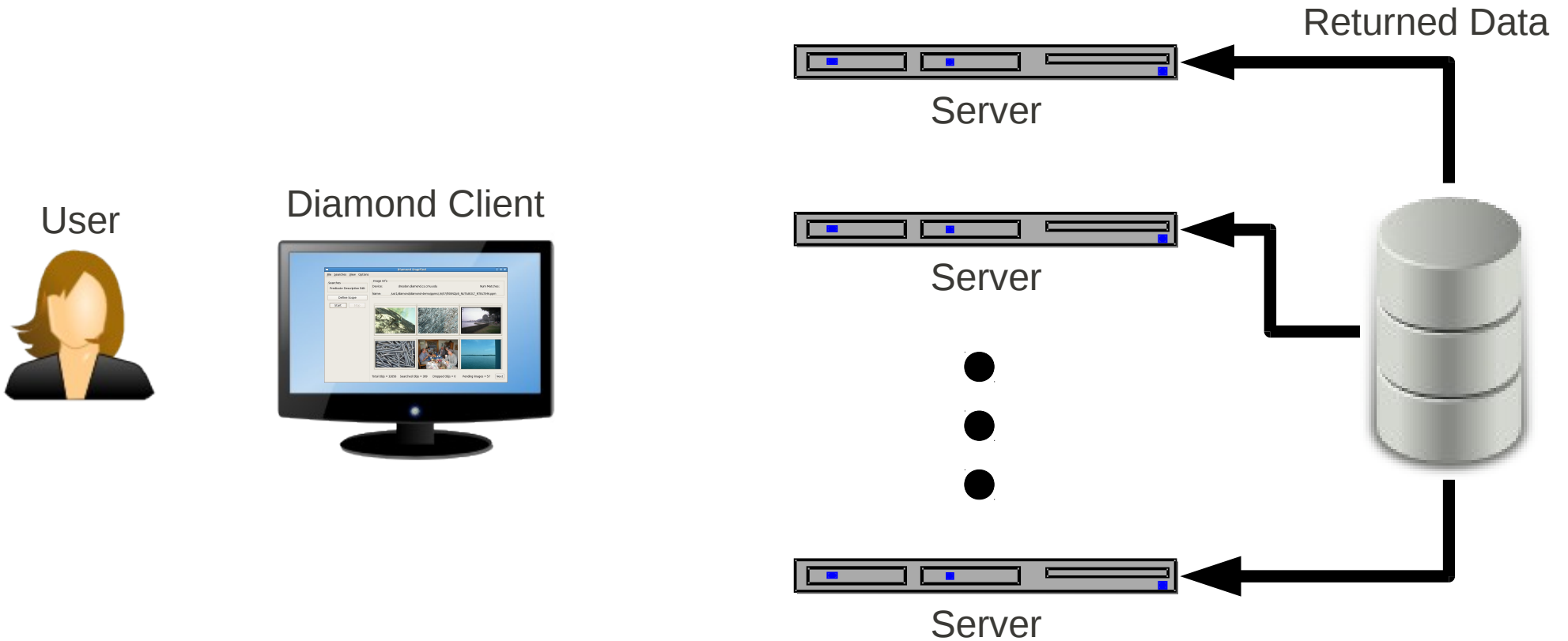
Discard-Based Search: Diamond



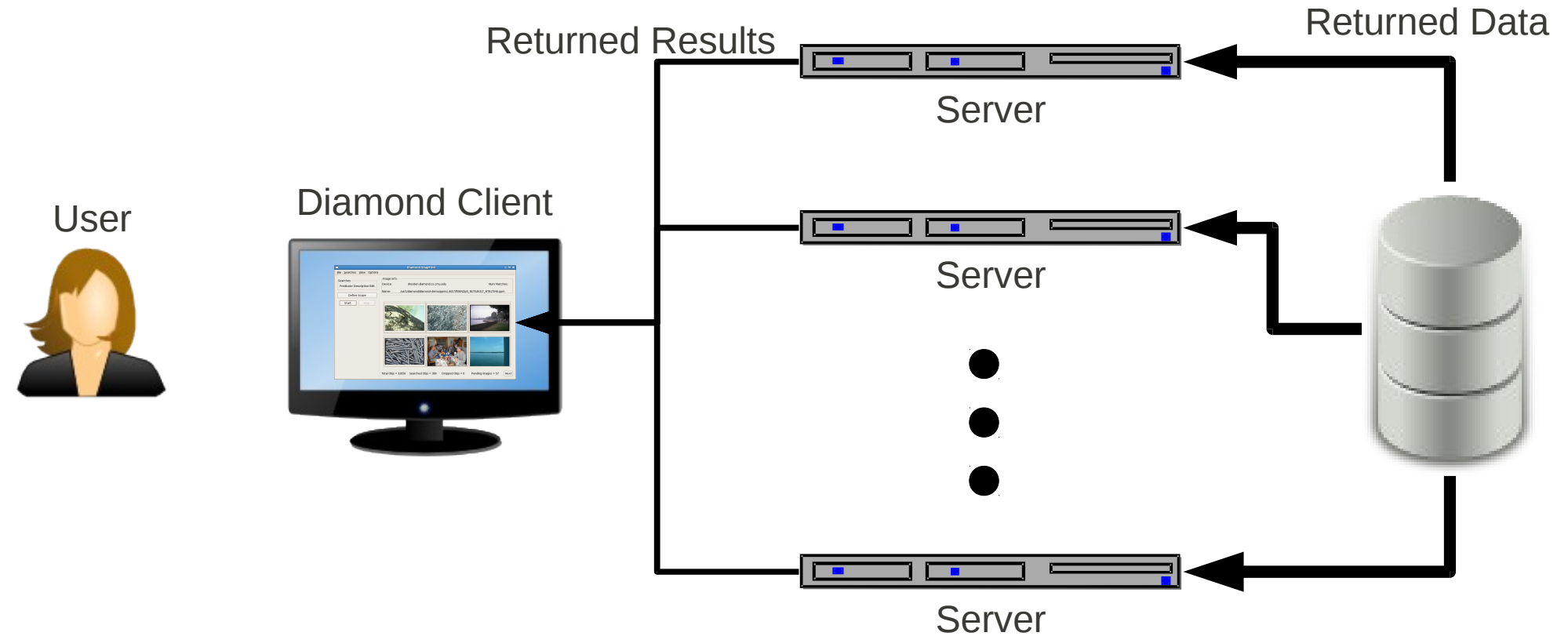
Discard-Based Search: Diamond



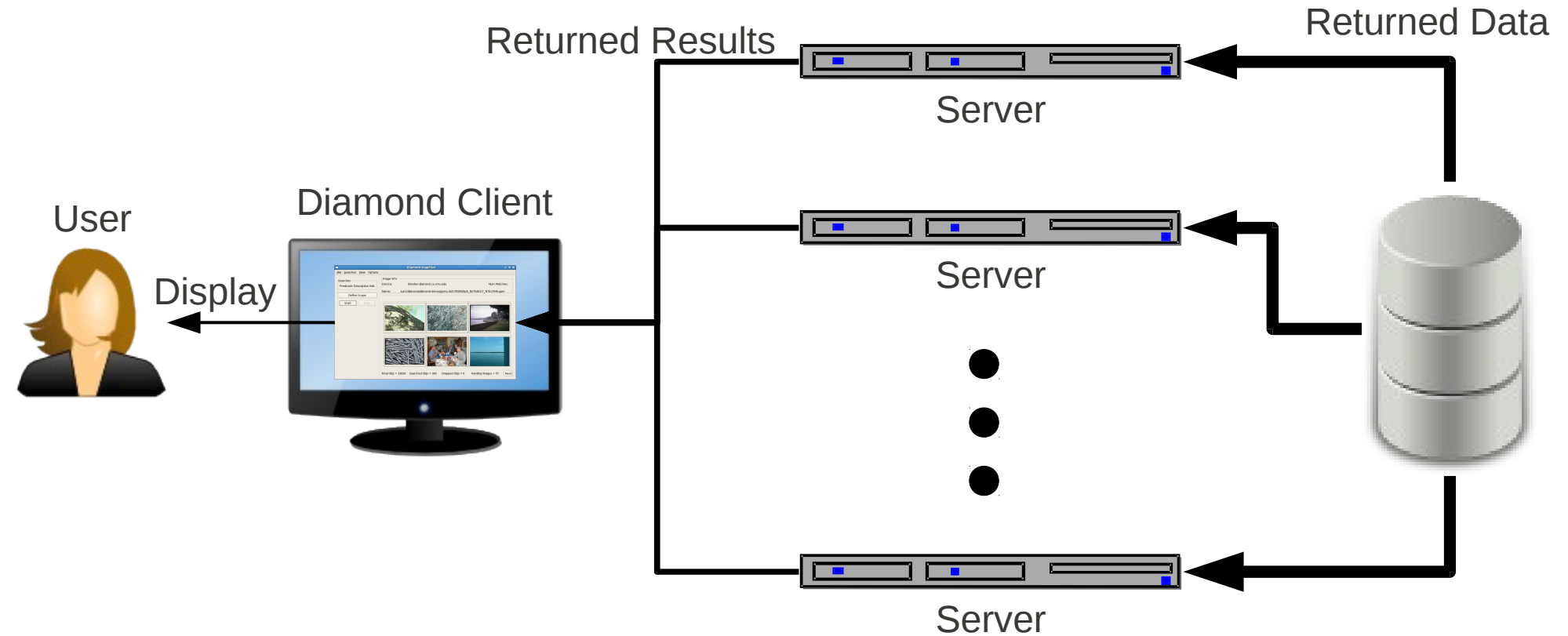
Discard-Based Search: Diamond



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

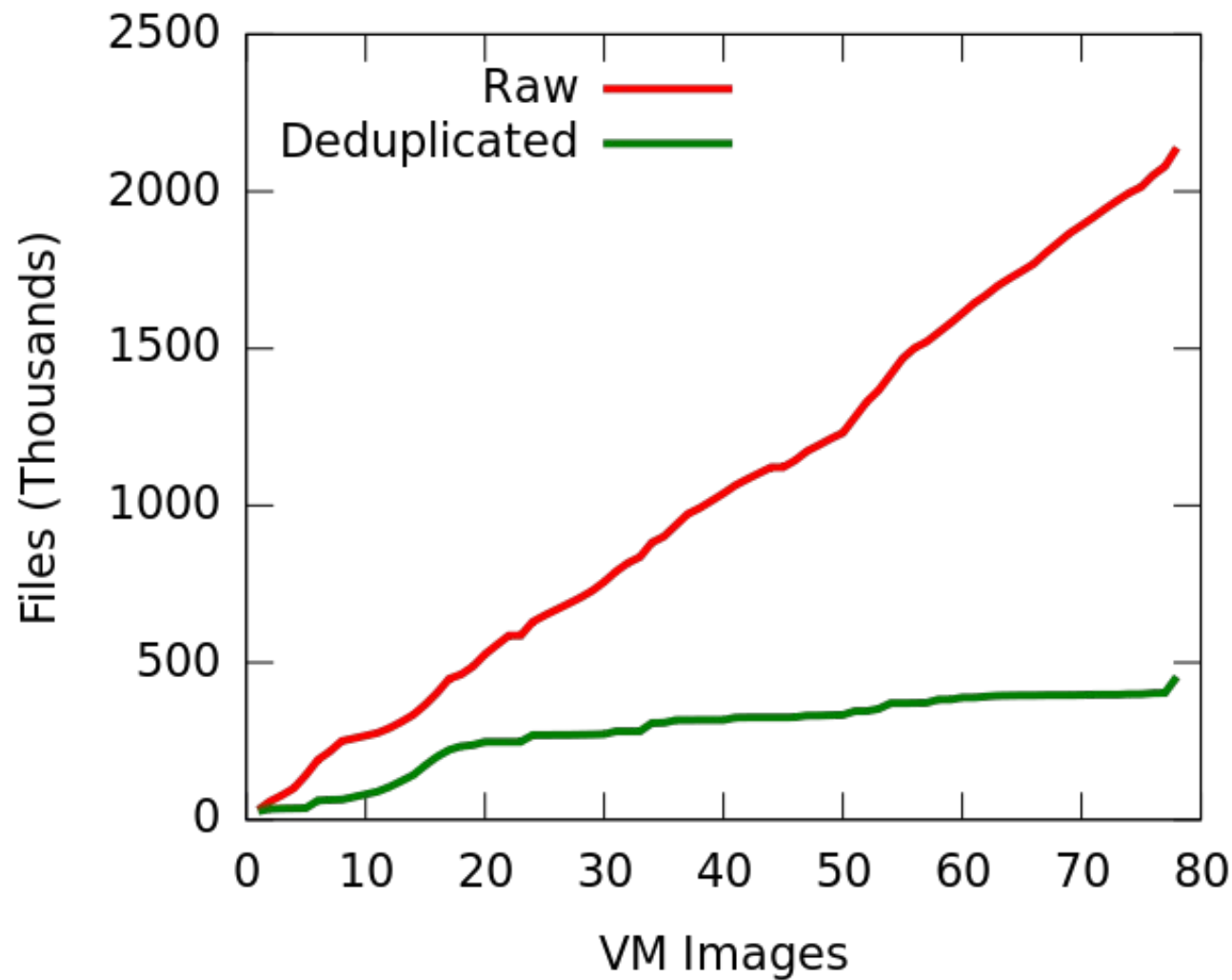
VM Growth

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

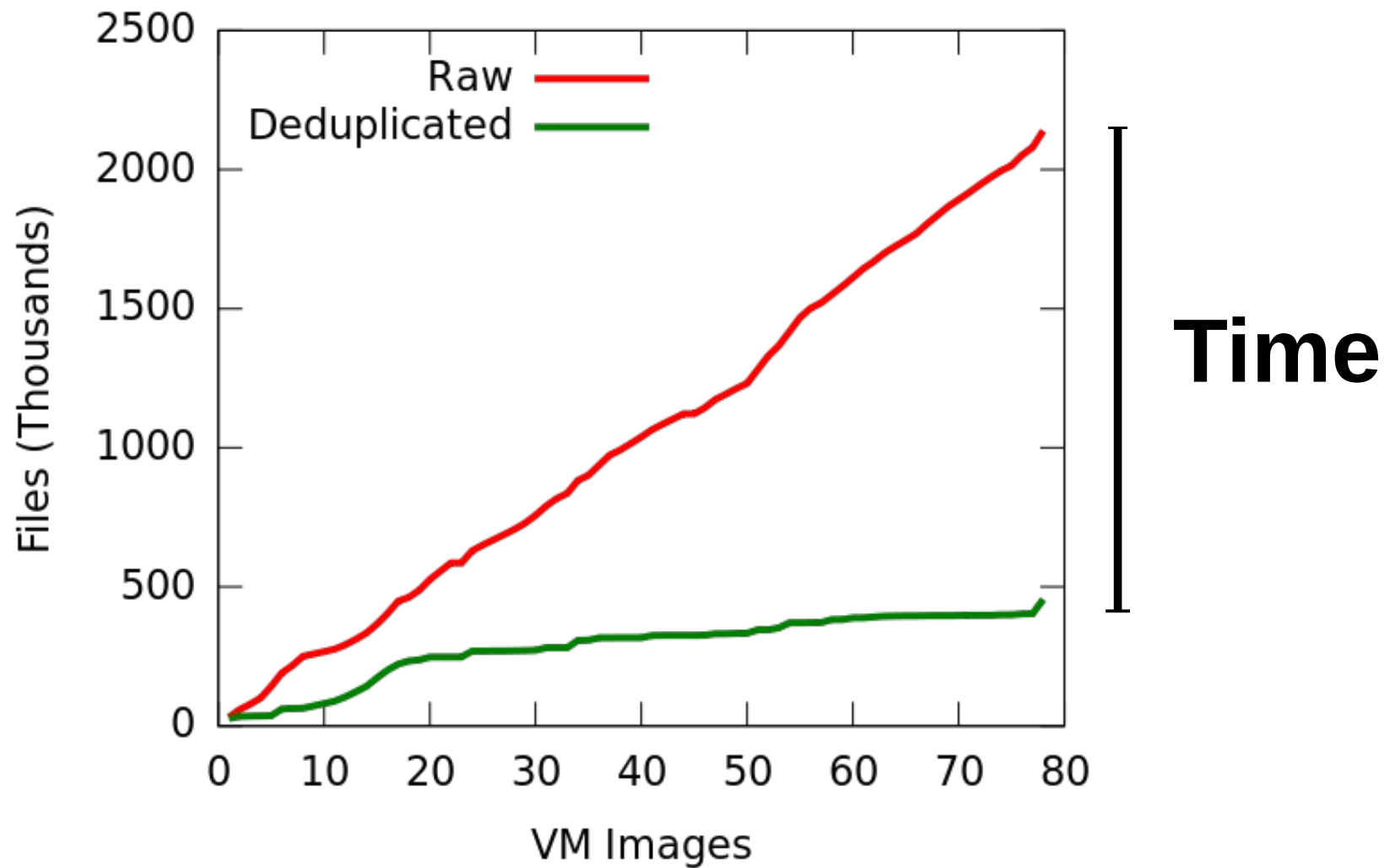
VM Search Solved?

- 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

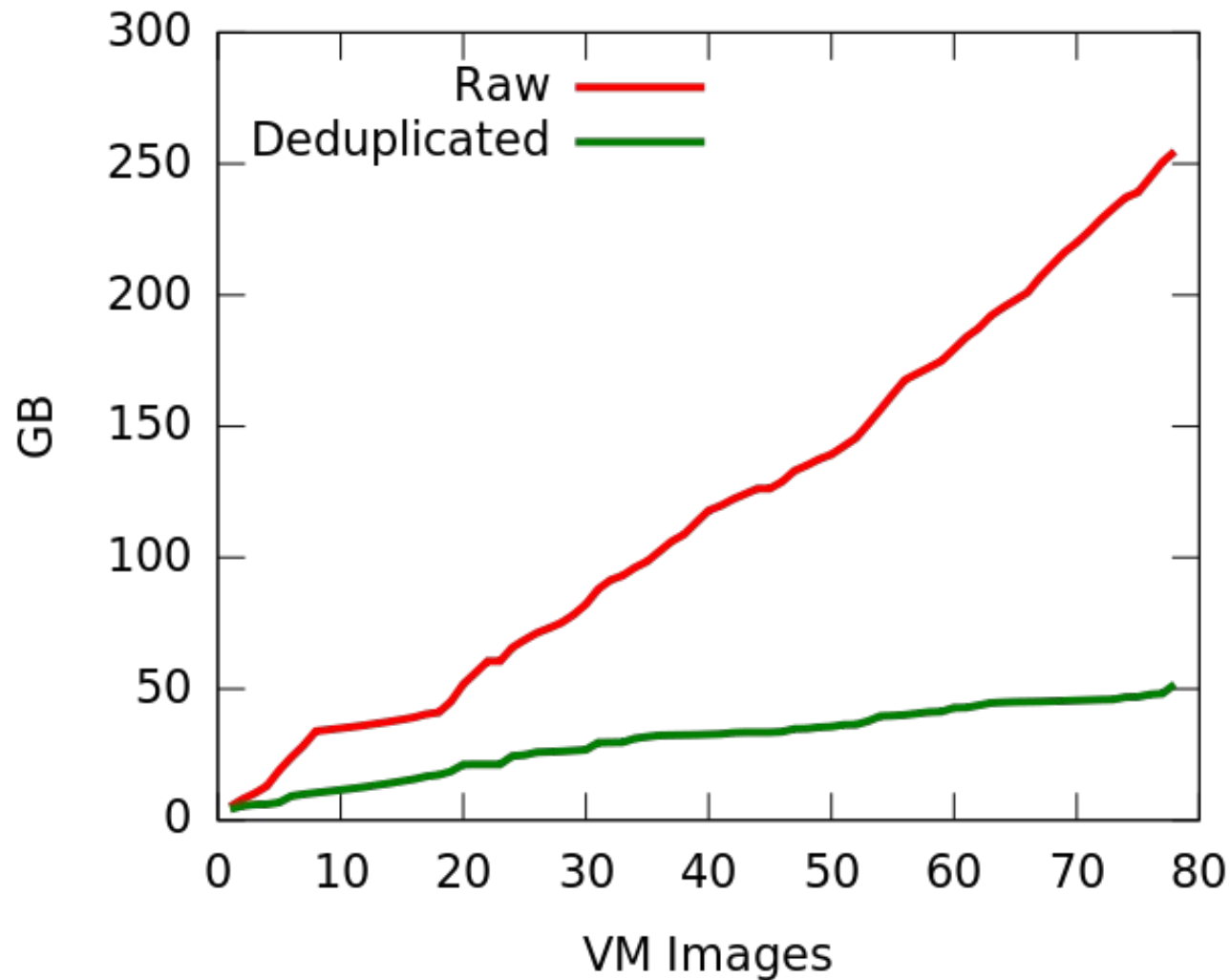
Importance of Deduplication: Files



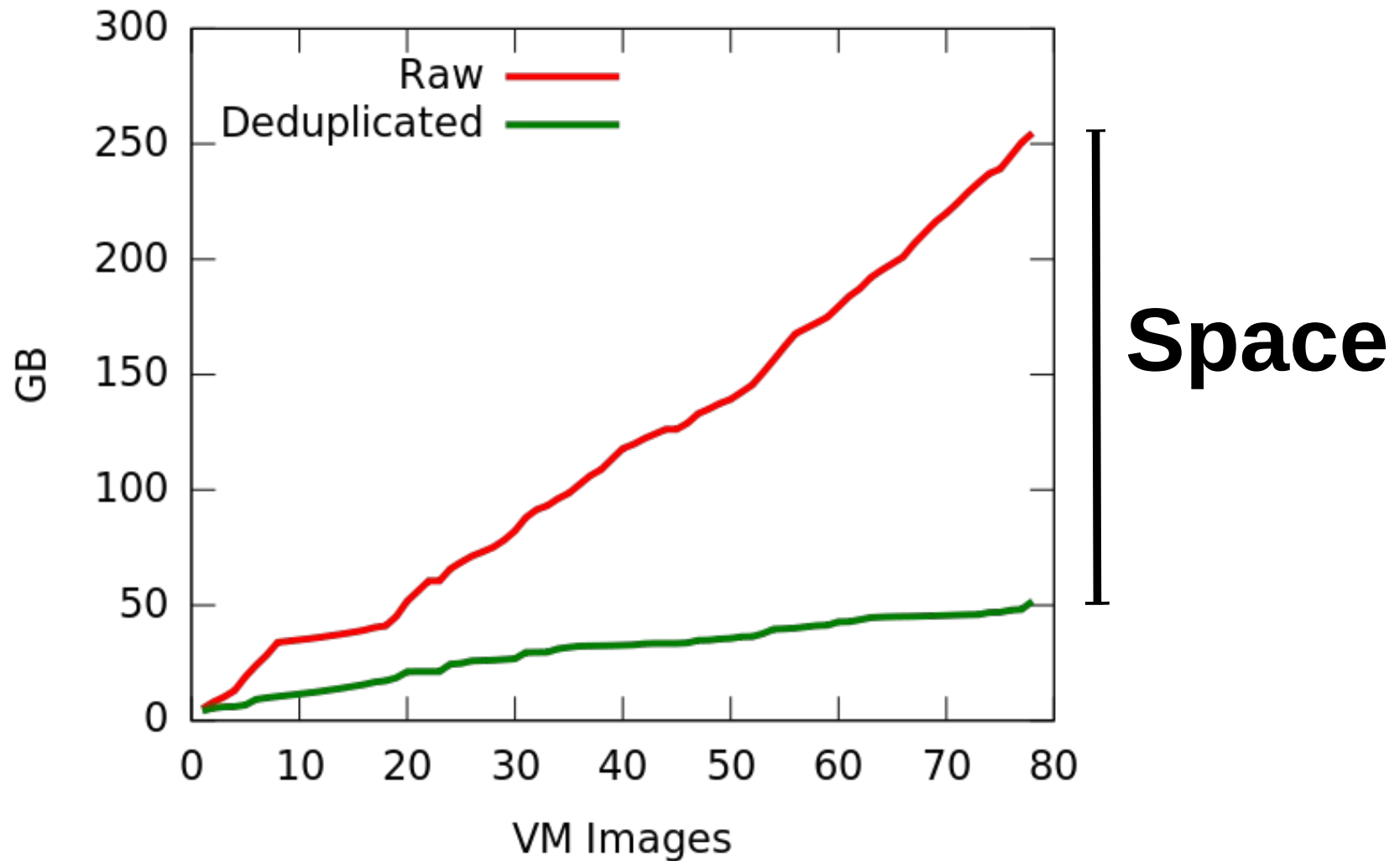
Importance of Deduplication: Files



Importance of Deduplication: Bytes



Importance of Deduplication: Bytes



How to **Deduplicate** VM's?

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

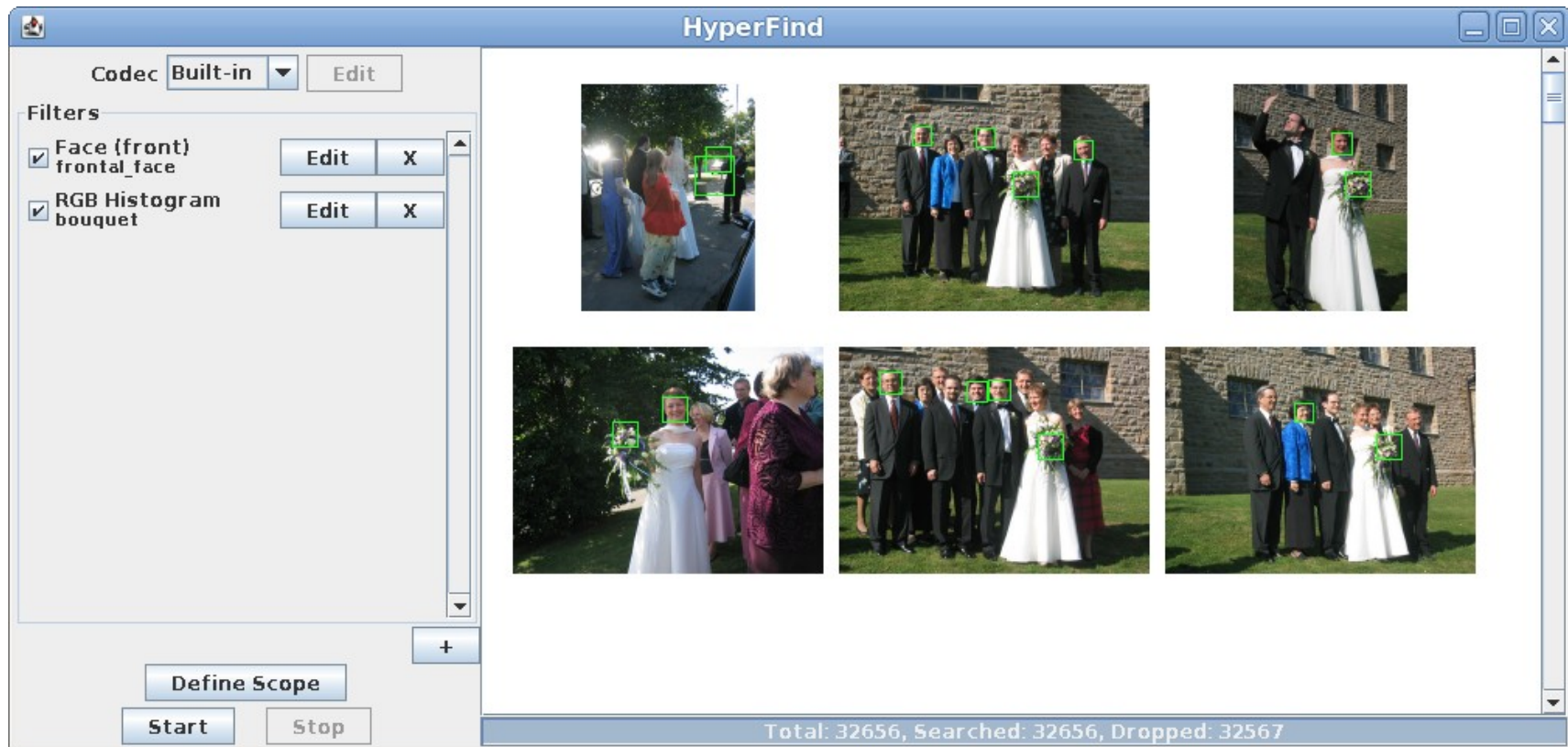
Deduplication Solution: **Mirage**

- 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

The screenshot displays the HyperFind application interface, which is used for source code search. The interface is divided into two main panels.

Left Panel (Filters):

- Codec:** Null codec (with an Edit button).
- Filters:**
 - ☒ W-Shingling filter (with Edit and X buttons).
 - Edit filter (delft):** A sub-panel with a Close button.
- Search:**
 - Filter Name: filter
- Shingling Parameters:**
 - Min Similarity: 0.30 (with a slider and up/down arrows).
 - Shingle size: 4 (with up/down arrows).
- Search Fragment:** foo = fork()
- Buttons:** Define Scope, Start, Stop, and a plus sign (+).

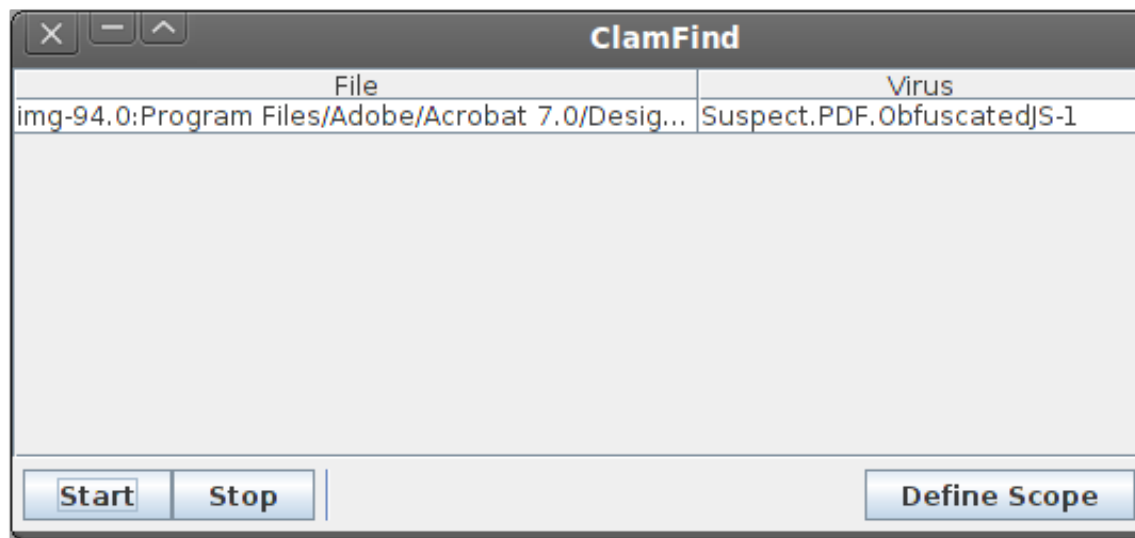
Right Panel (Results):

- File List:**
 - img-147.0:./usr/src/linux-2.6.27.43/scripts/basic/docproc.c
 - img-146.0:./usr/src/linux-2.5.75/arch/um/kernel/tt/ptproxy/proxy.c
 - img-147.0:./usr/src/linux-2.6.27.43/scripts/basic/docproc.c
- Code View:** Displays the source code for the selected file (img-147.0:./usr/src/linux-2.6.27.43/scripts/basic/docproc.c). The code includes a switch statement for `pid=fork()`.
- Table:** A table with 3 columns: Name, Size, and Value.

Name	Size	Value
Device-Name	6	biela
Display-Name	60	img-147.0:./usr/src/linux-2.6.27.43/scrip...
FIL_TIME.time	8	0
FIL_TIME_RGB.time	8	0
FIL_TIME_THUMBNAI.ti...	8	0
FIL_TIME_z4da8e58cb9...	8	0
ObjectID	84	http://fram.isr.cs.cmu.edu:5873/mirage/...
cols.int	4	32
rows.int	4	32

Total: 40618, Searched: 40571, Dropped: 40562

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?

