

Application Buffer-Cache Management for Performance: Running the World's Largest MRTG

LISA '07, November 14, 2007



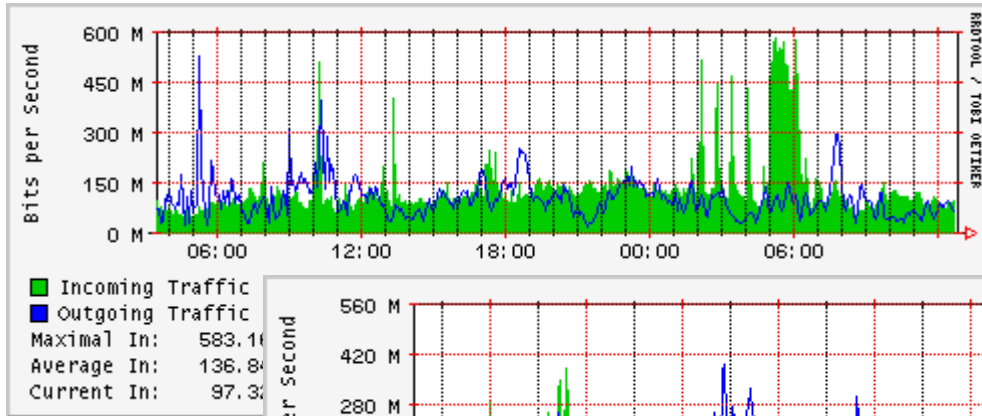
David Plonka, Archit Gupta, and Dale Carder
{plonka,archit}@cs.wisc.edu, dwcarder@doit.wisc.edu

Outline

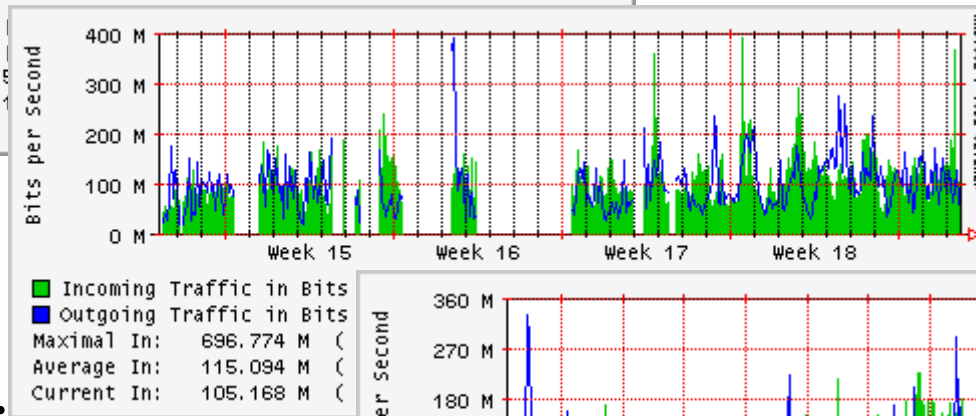
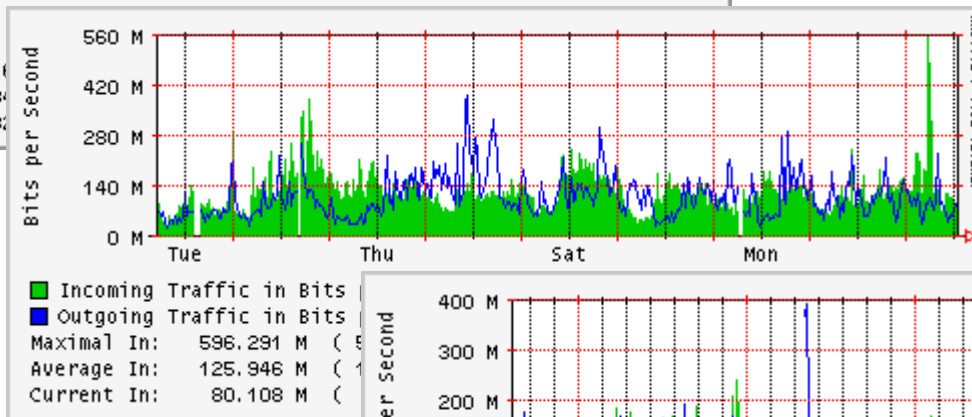
- Background on MRTG and RRDTool
- Problem & Motivation
- Investigative Tools & Methodology
- Approach 1: Application-level Buffering
- Approach 2: Application Advice
- Scalability
- Analysis: Model & Simulation
- Summary and Contributions
- “Questions?”

What is MRTG?

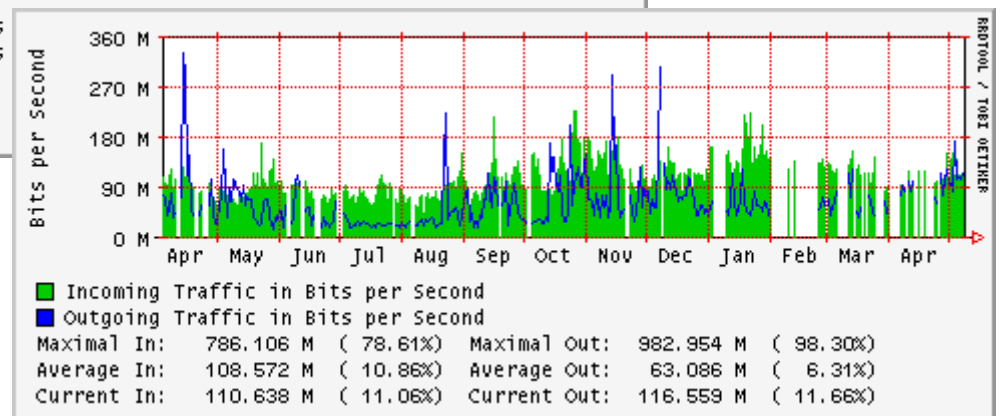
Daily, 5 min averages



Weekly, 30 min ave.

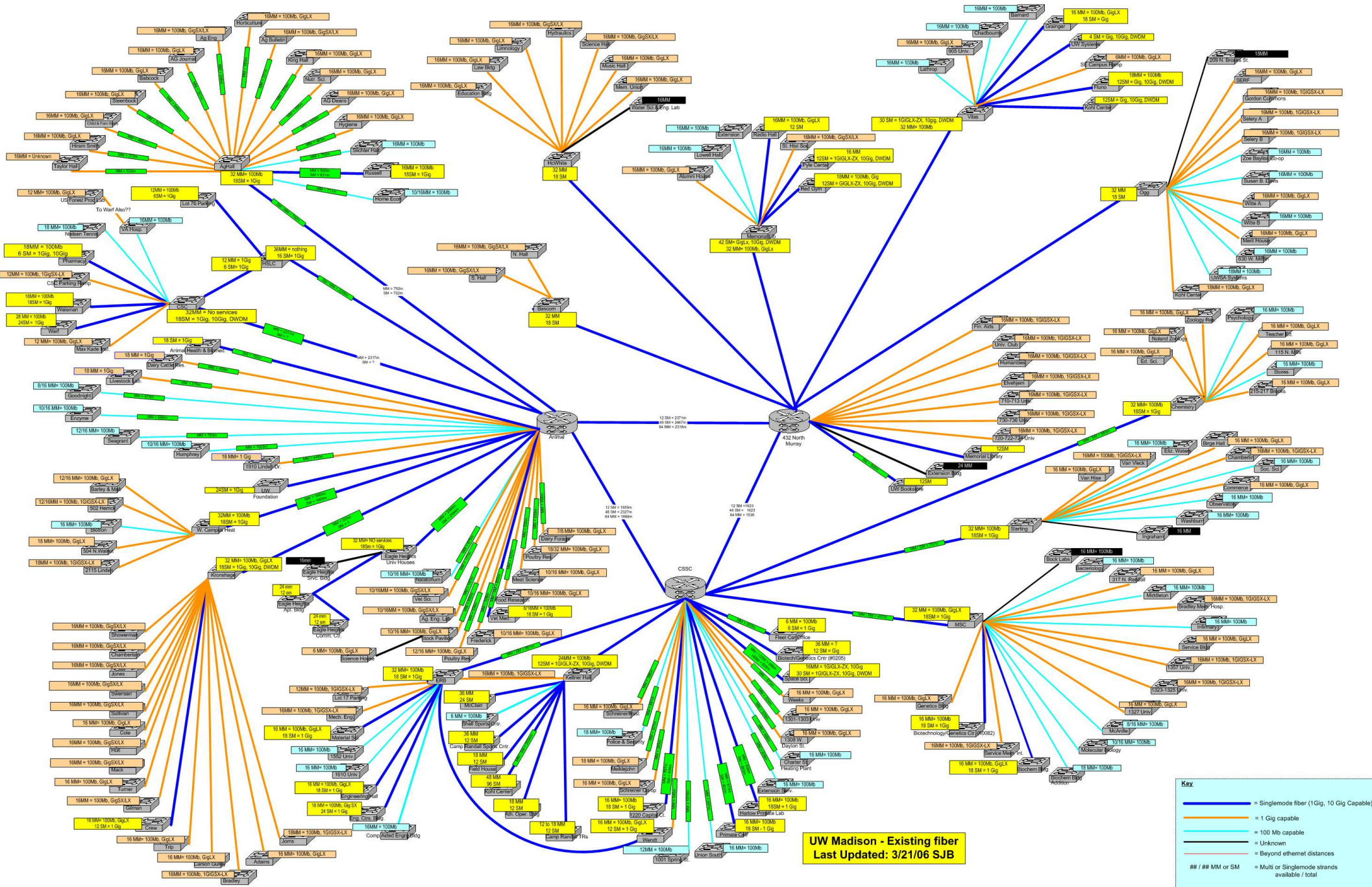


Monthly, 2 hour ave.

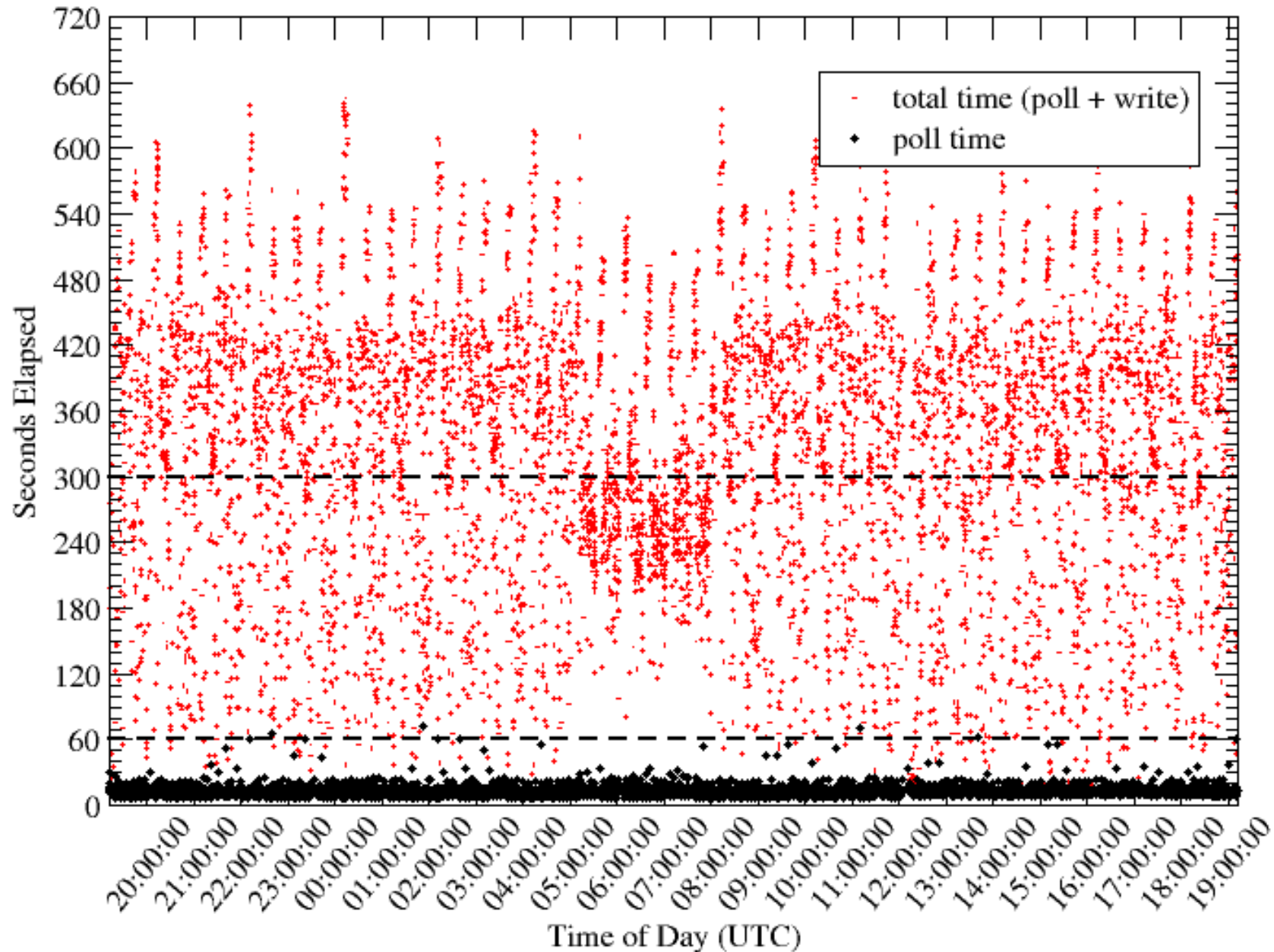


Yearly, 1day averages

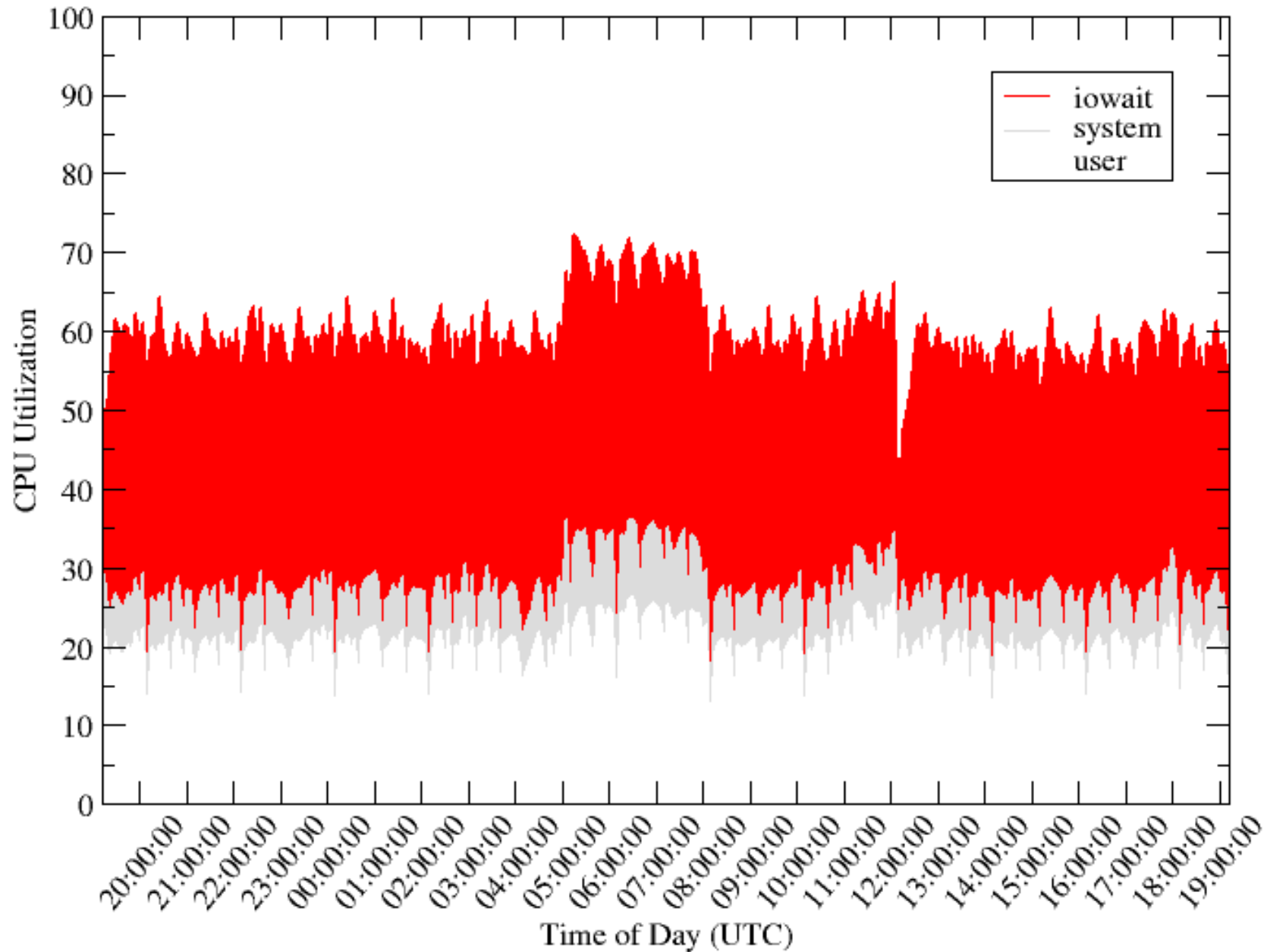
Our Network



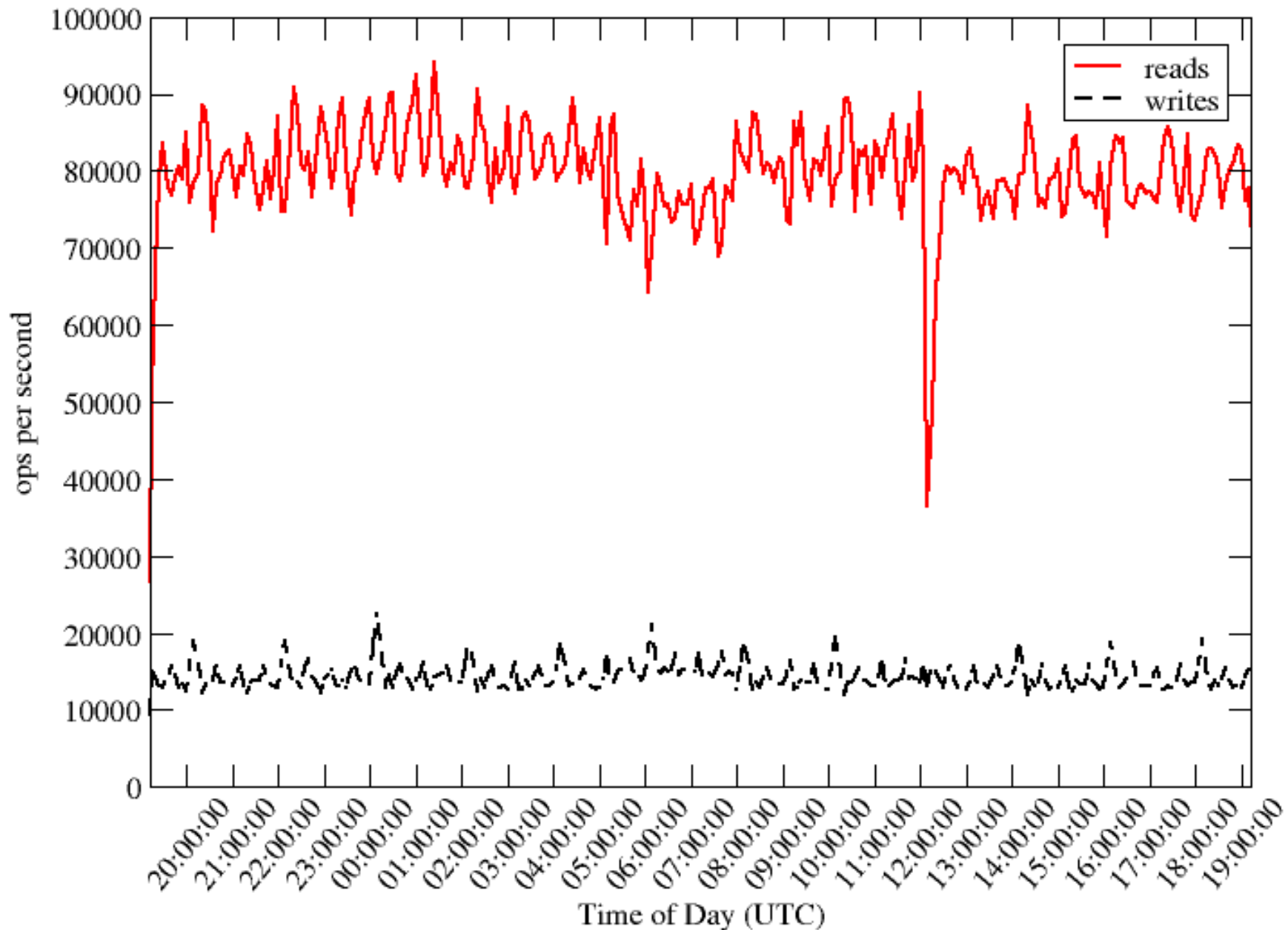
What's the Problem?



What's the Problem?



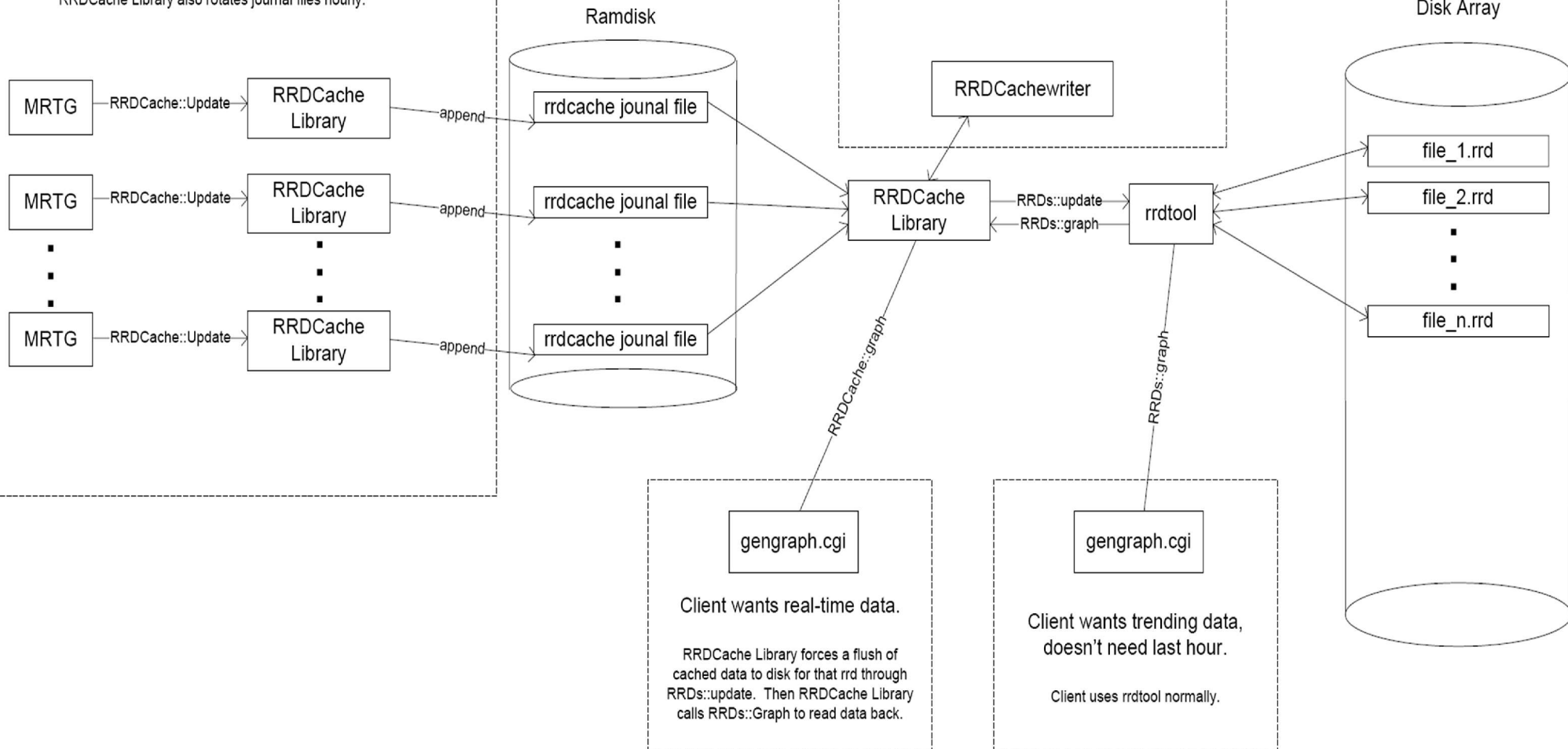
What's the Problem?



RRDCache: App-level Buffering

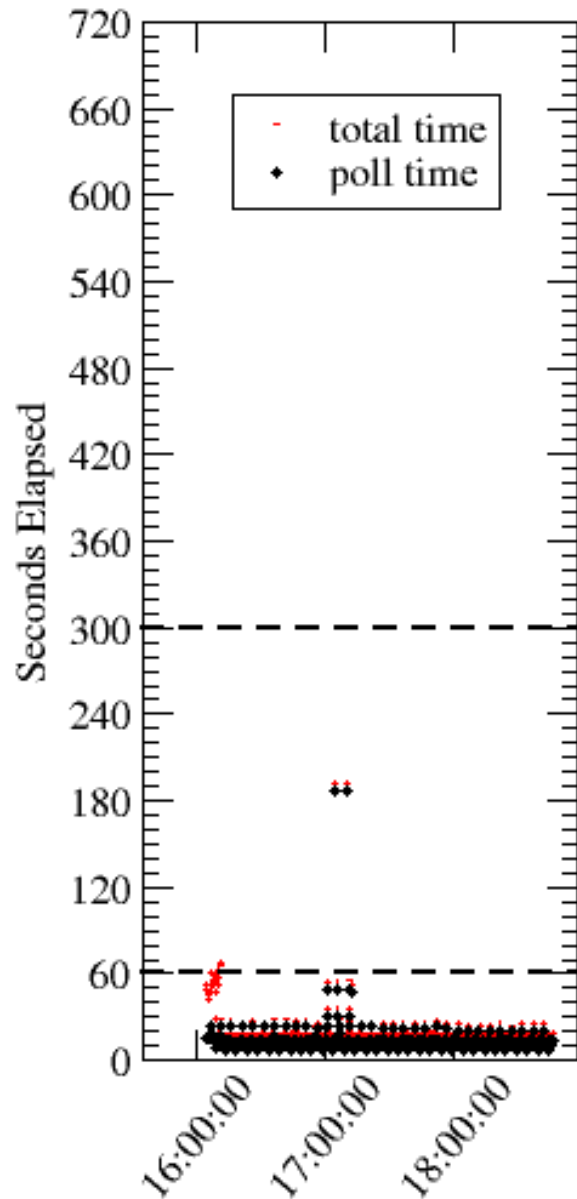
Every 5 minutes, append time&data to journal.

RRDCache Library also rotates journal files hourly.

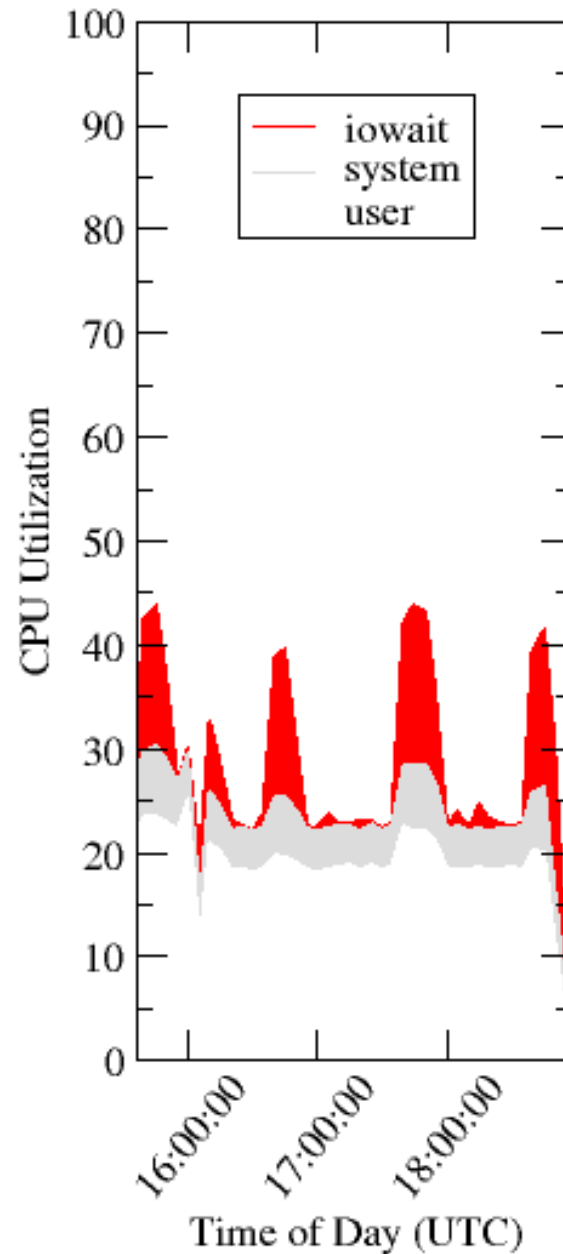


RRDCache Performance: **Good!**

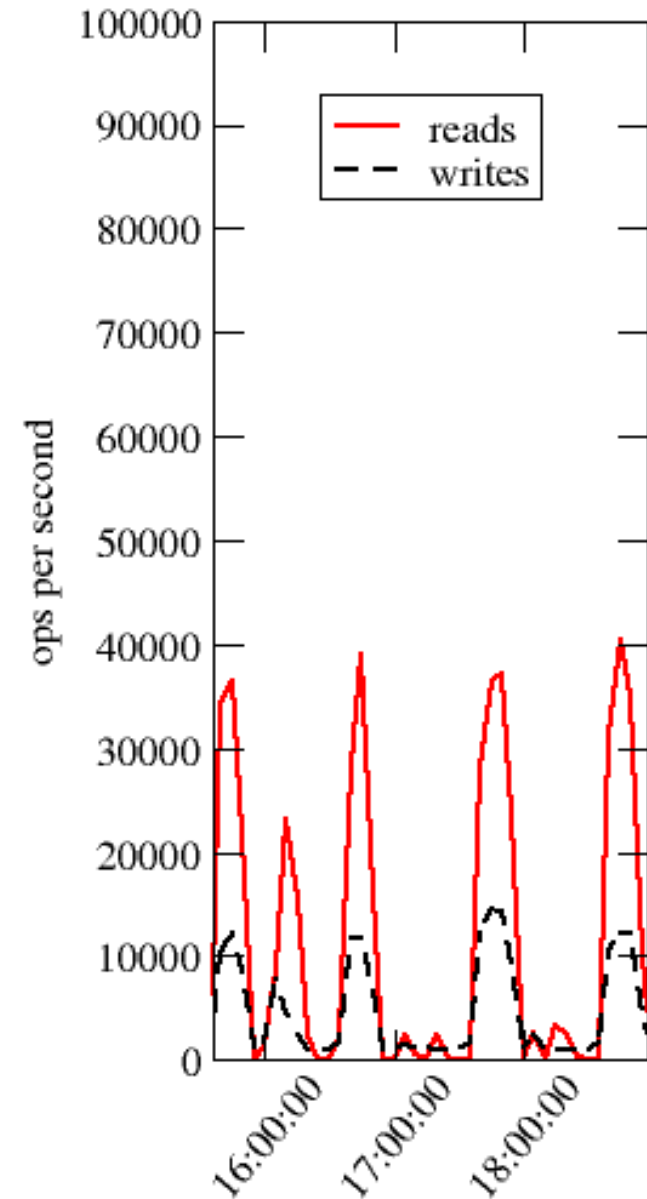
MRTG



CPU



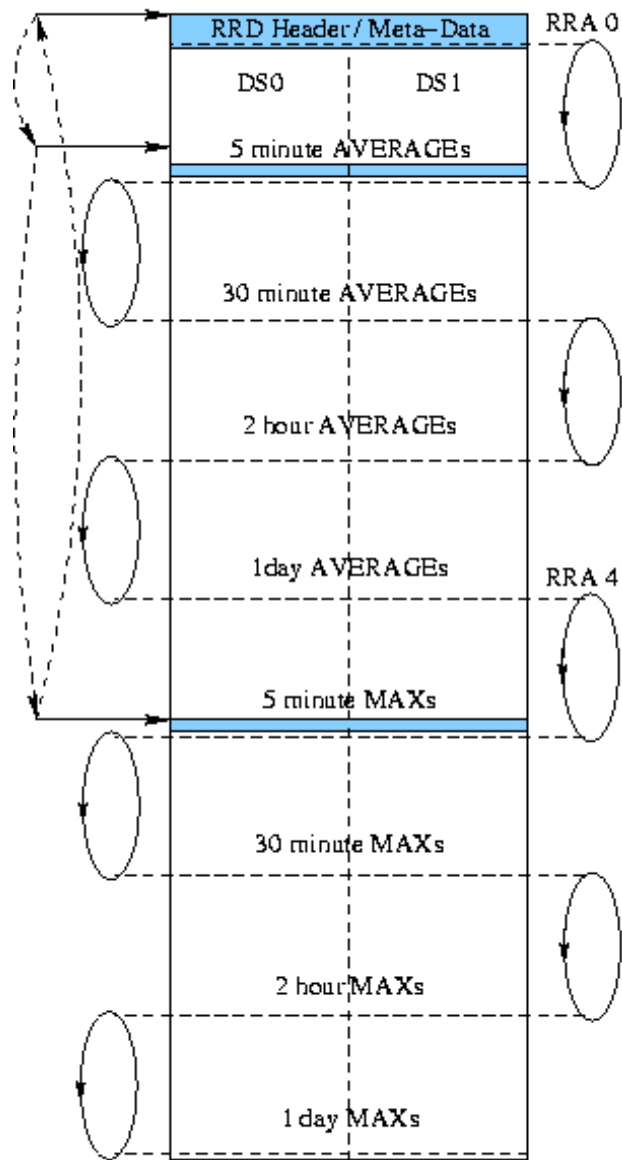
Disk I/O



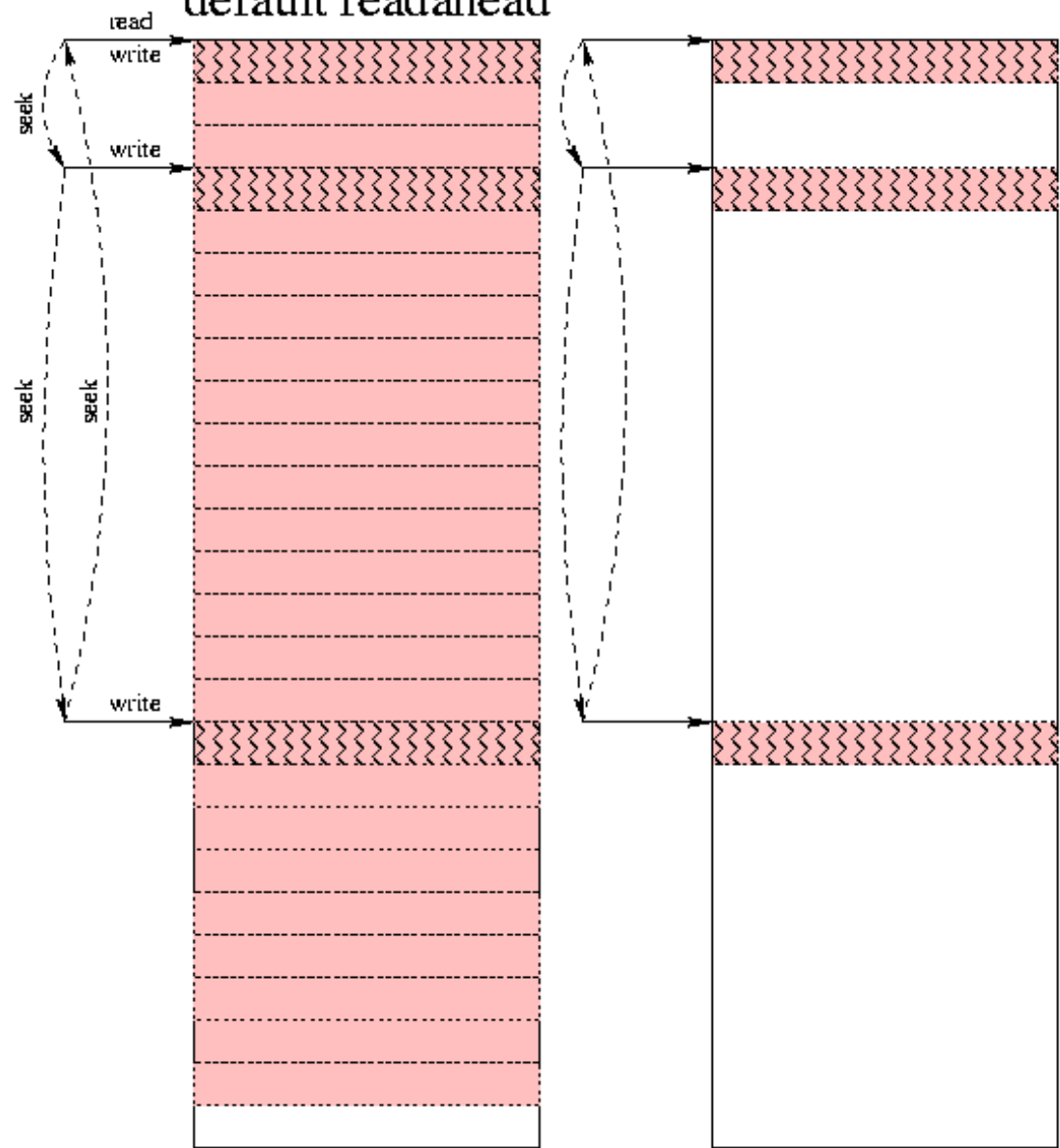
Methodology & Tools

- **sar**
 - System activity reports (CPU, disk, etc.)
 - Can visualize data with new ksar tool
- **fincore**
 - A user command that exposes the cache “footprint” of a set of files; shows which file blocks are in core
- **fadvise**
 - A user command that can forcibly evict a file's pages from buffer-cache
 - Invaluable for controlled experimentation




Typical RRD File



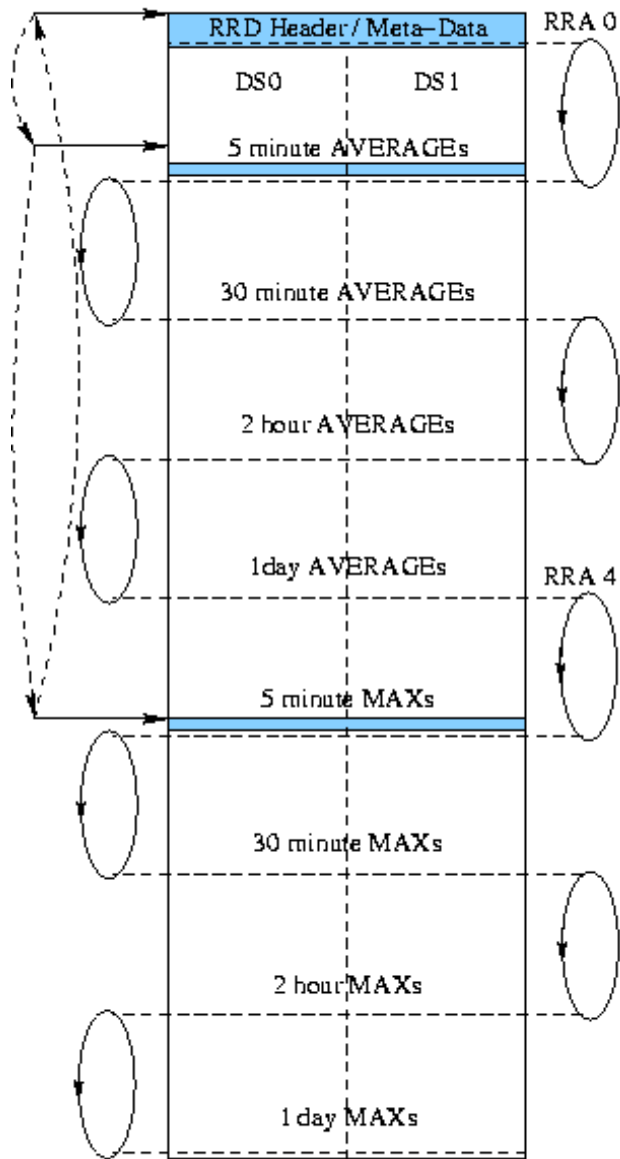
Linux 2.6.9 default readahead



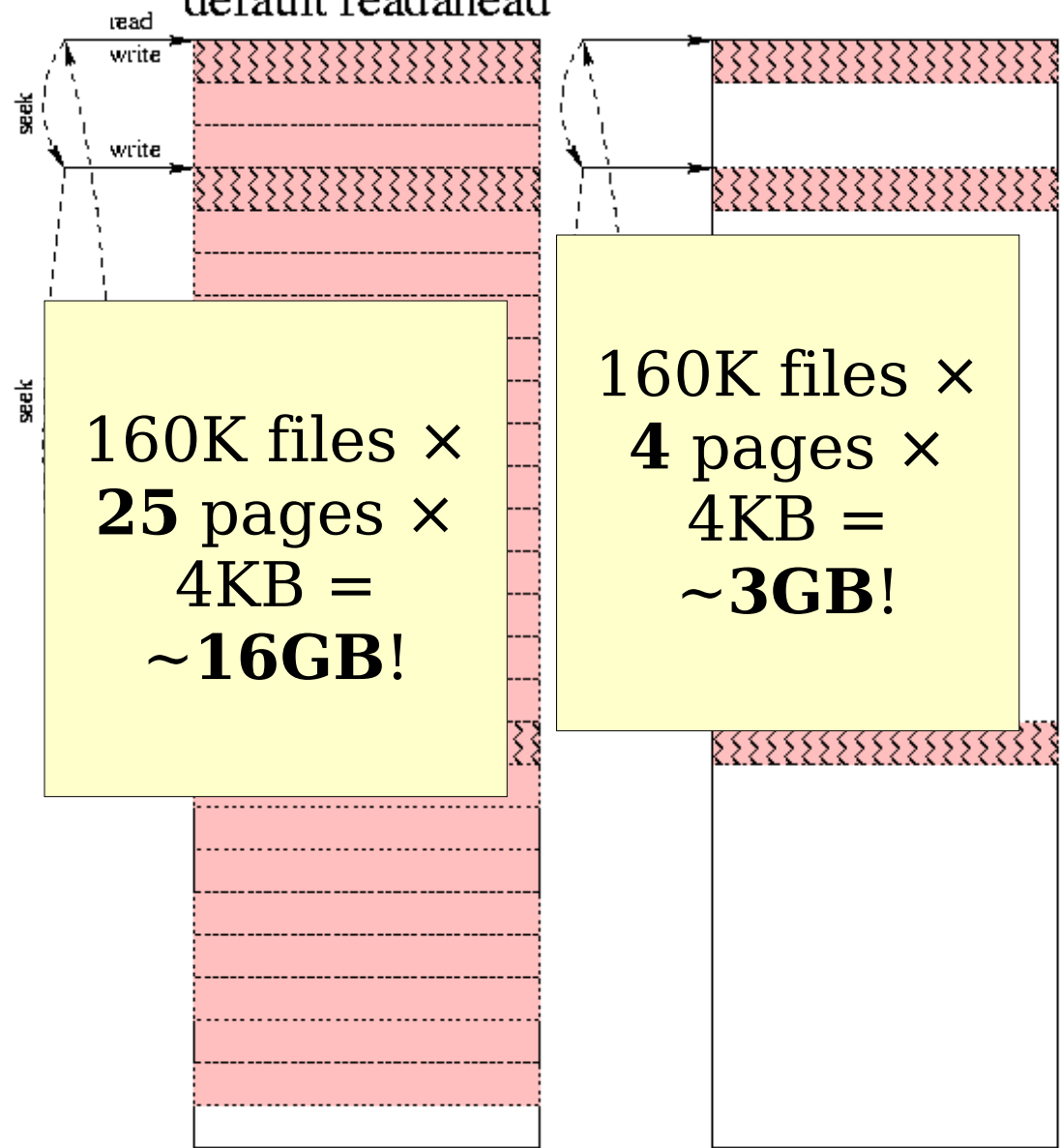
LEGEND

-  Actual File Update Area
-  Unnecessarily Read Block
-  "Hot" Block

Typical RRD File



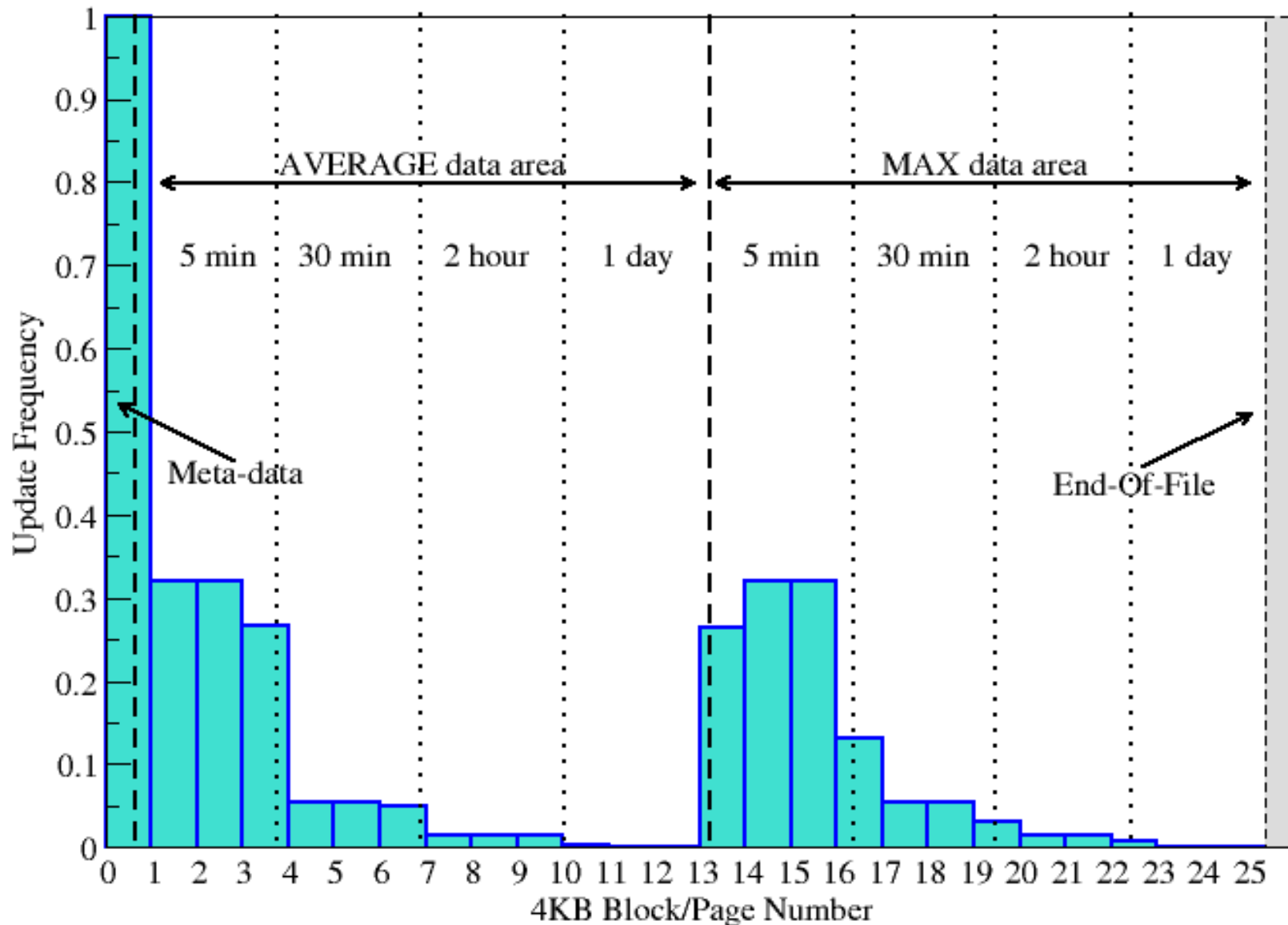
Linux 2.6.9 default readahead



LEGEND

- Actual File Update Area
- Unnecessarily Read Block
- "Hot" Block

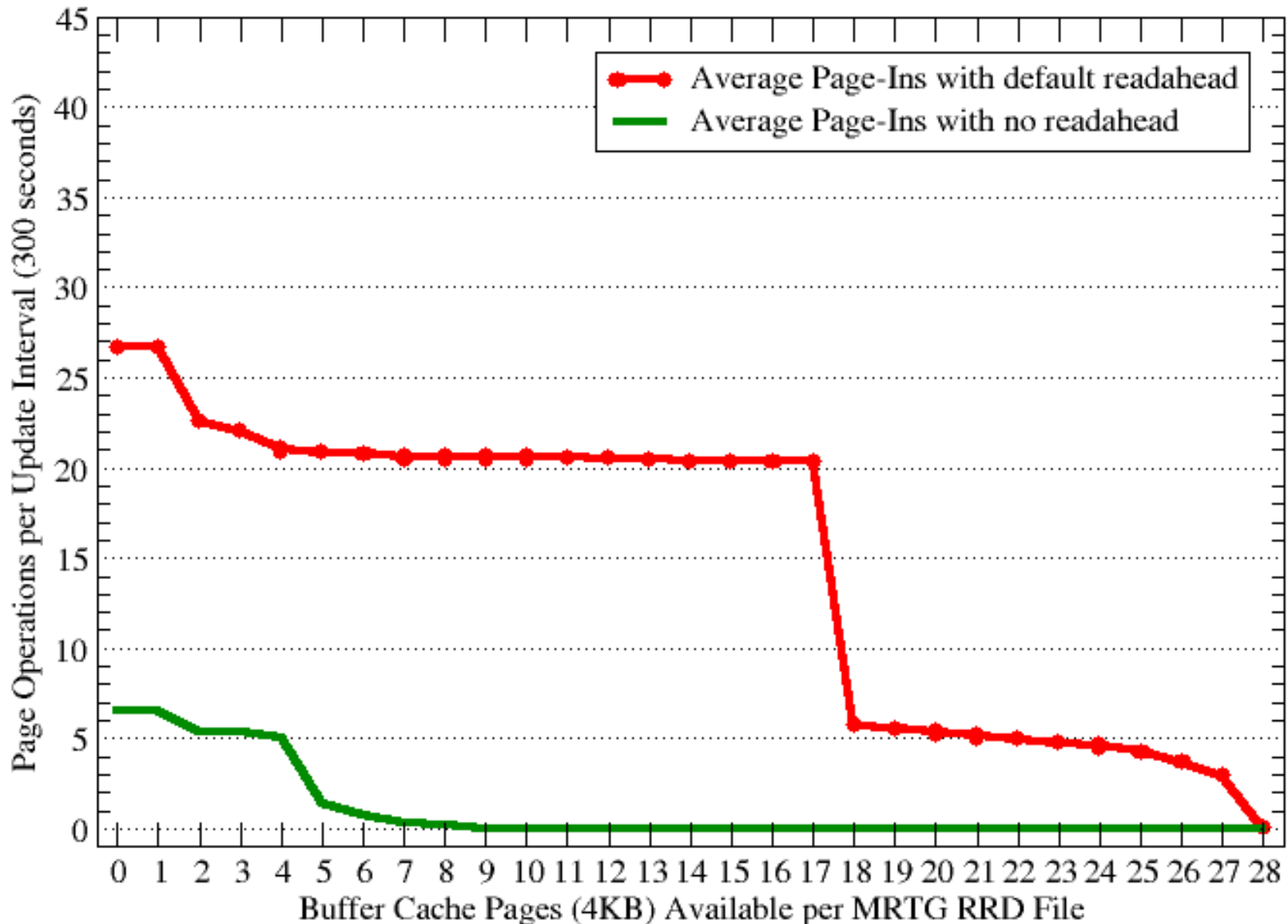
Simulation: Hot Pages



What's the Solution?

fadvise (RANDOM) ;

fdadvise (RANDOM) : Before & After

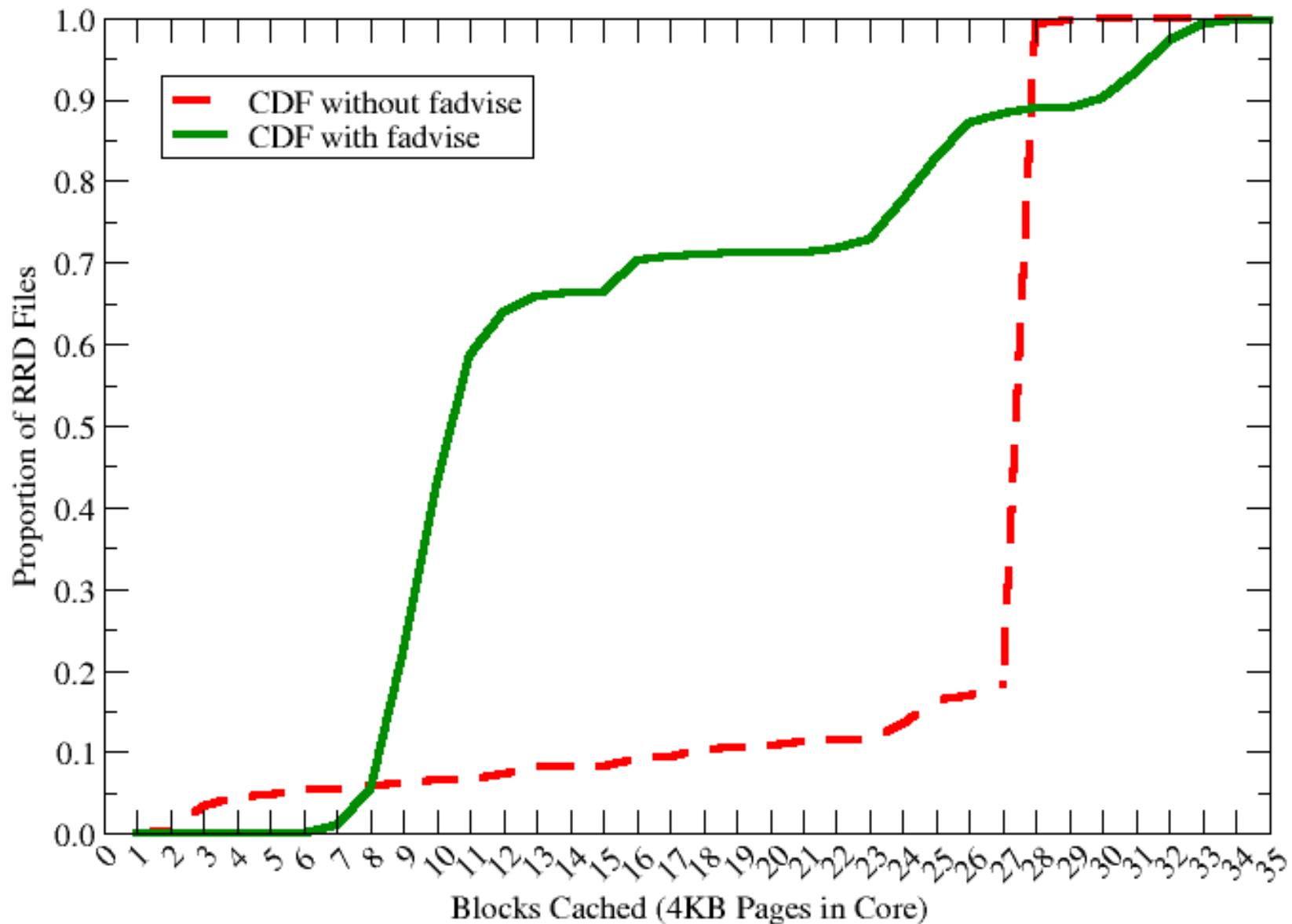


Analytical Model for Average RRD update Page Fault Rate

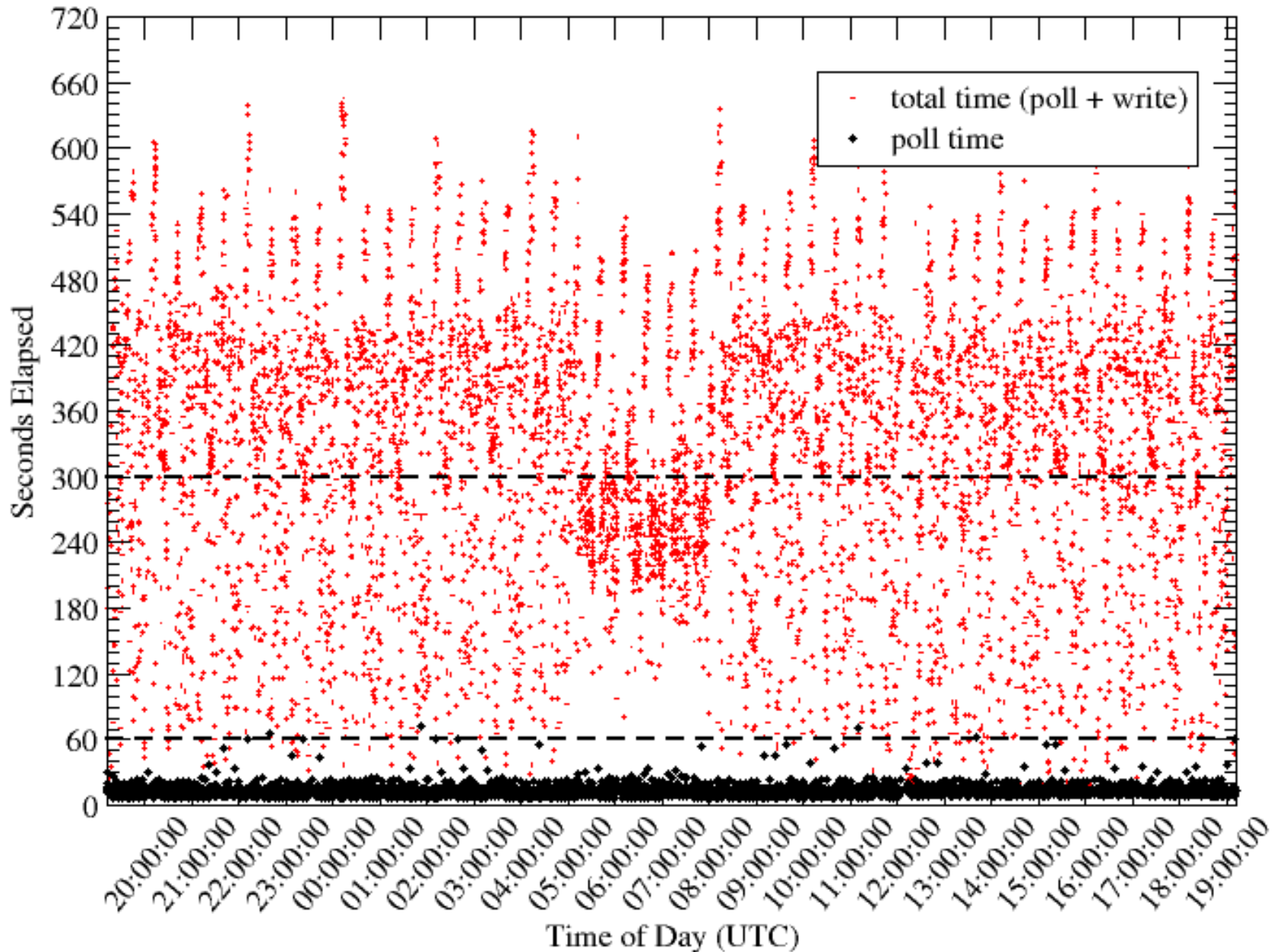
$$r = \frac{m}{t}$$

$$r = \frac{\delta \sum_{i=1}^n \frac{1}{x_i}}{\text{minimum}(uT, x_s T)}$$

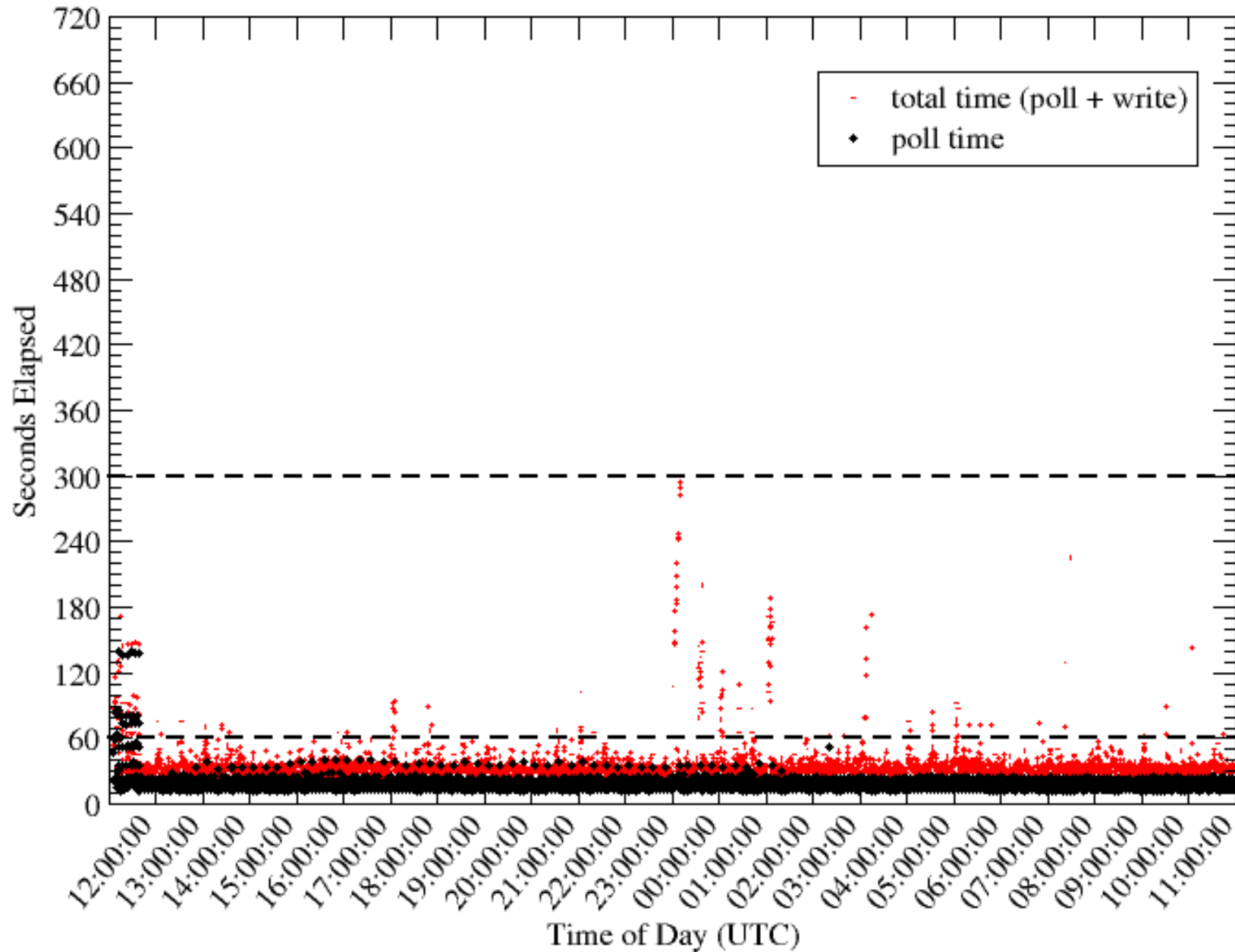
fadvice (RANDOM) Before & After



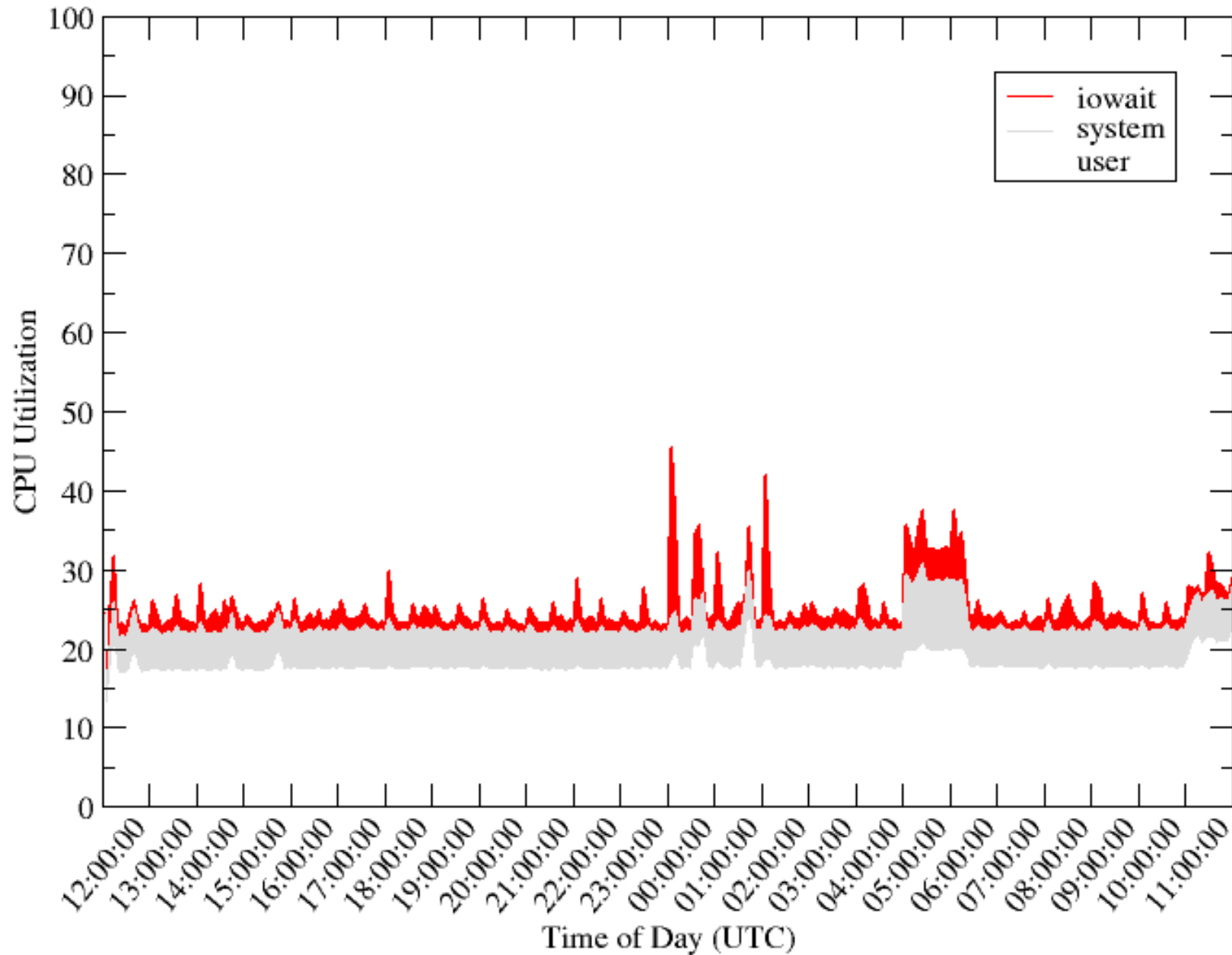
MRTG Performance: Before



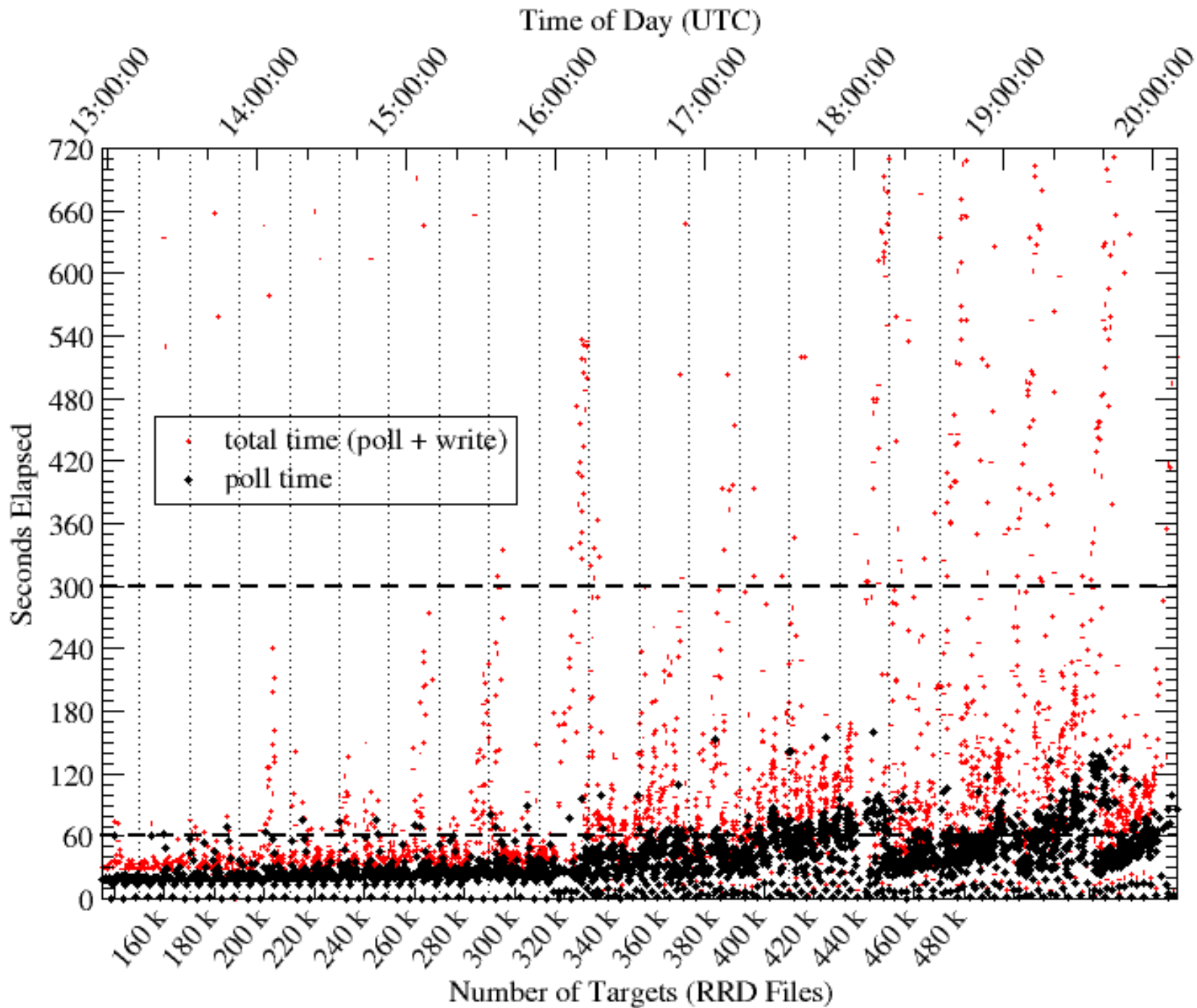
MRTG Performance: After. **Great!**



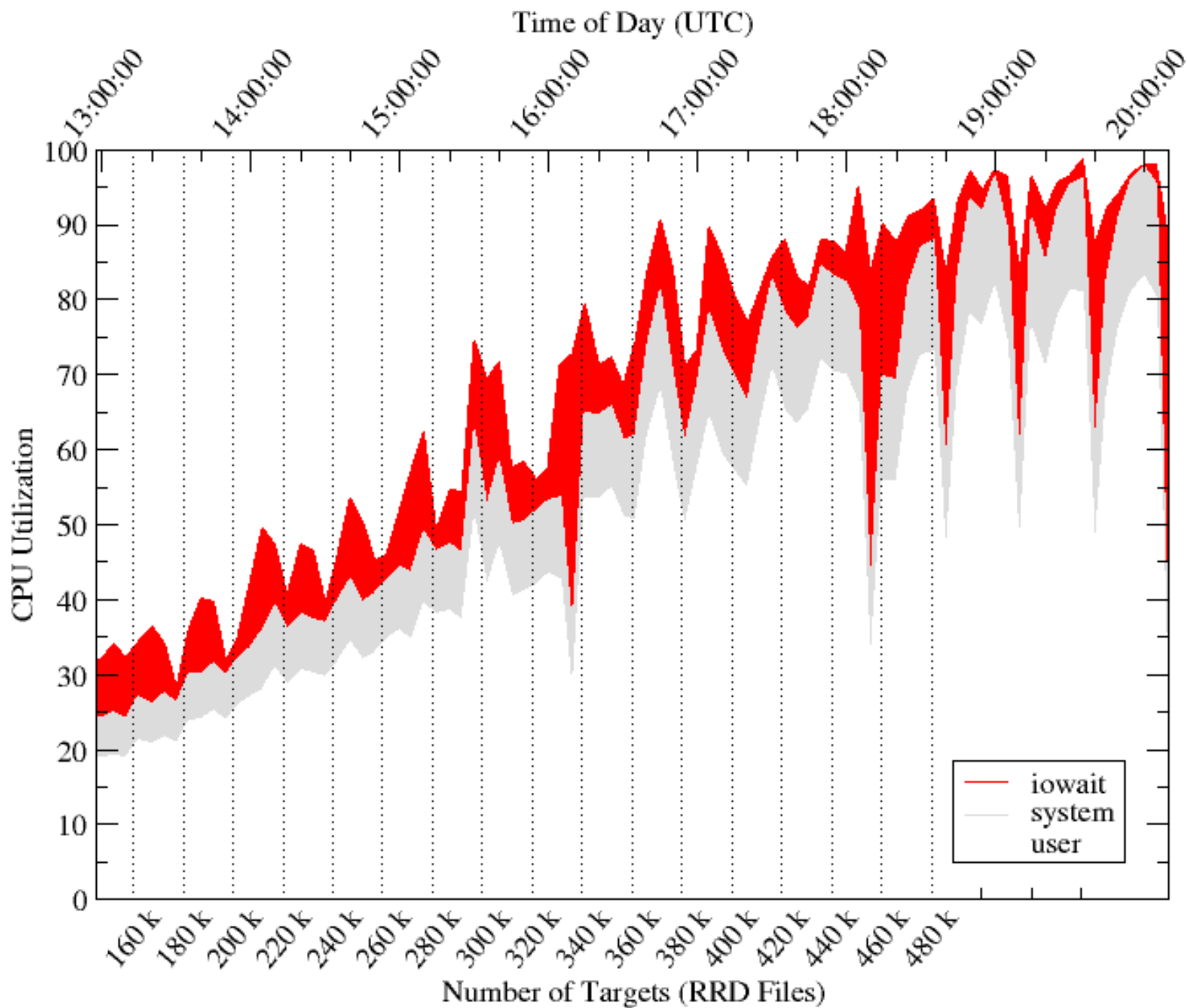
System CPU: After



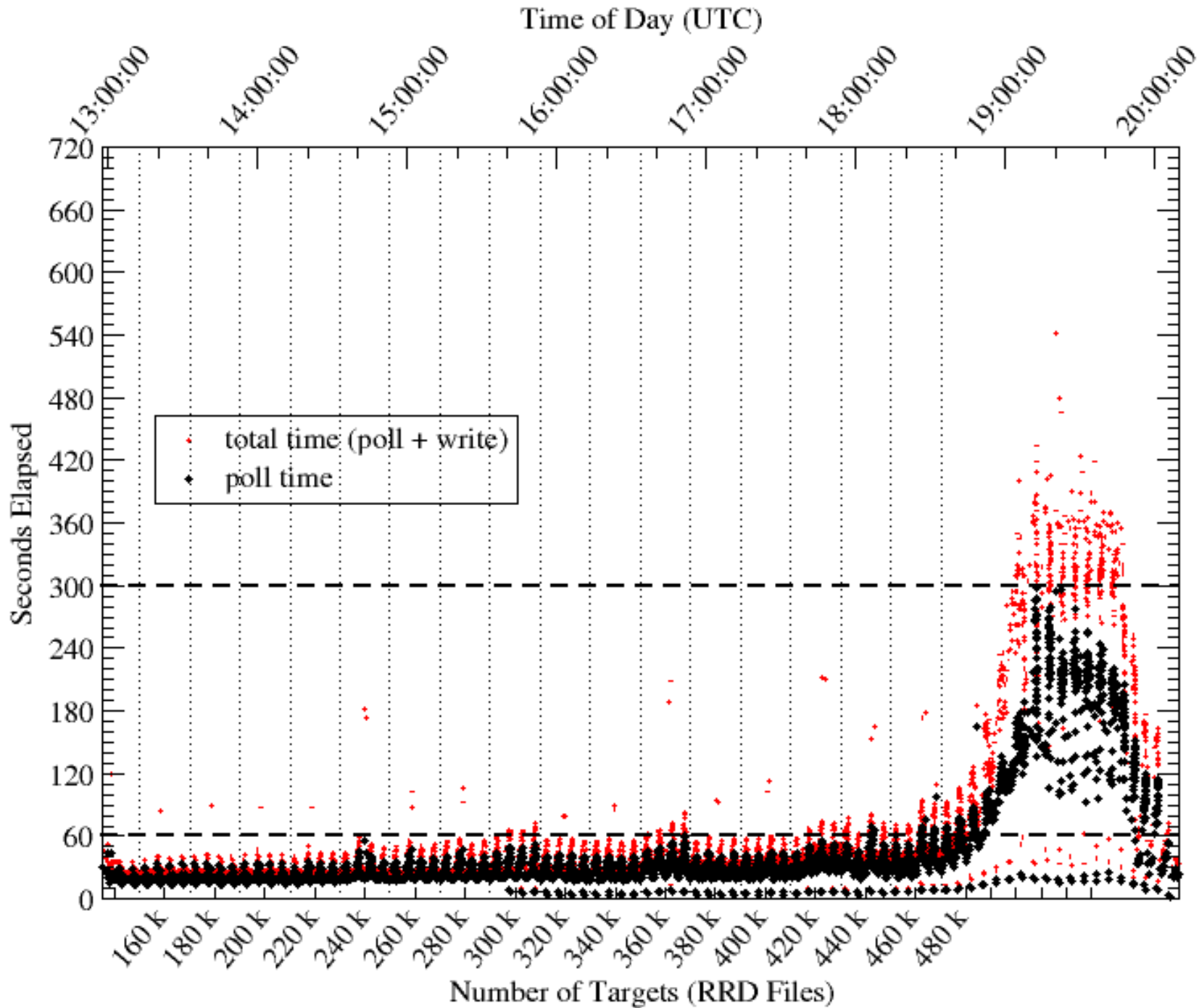
Scalability: `fadvice` (RANDOM)



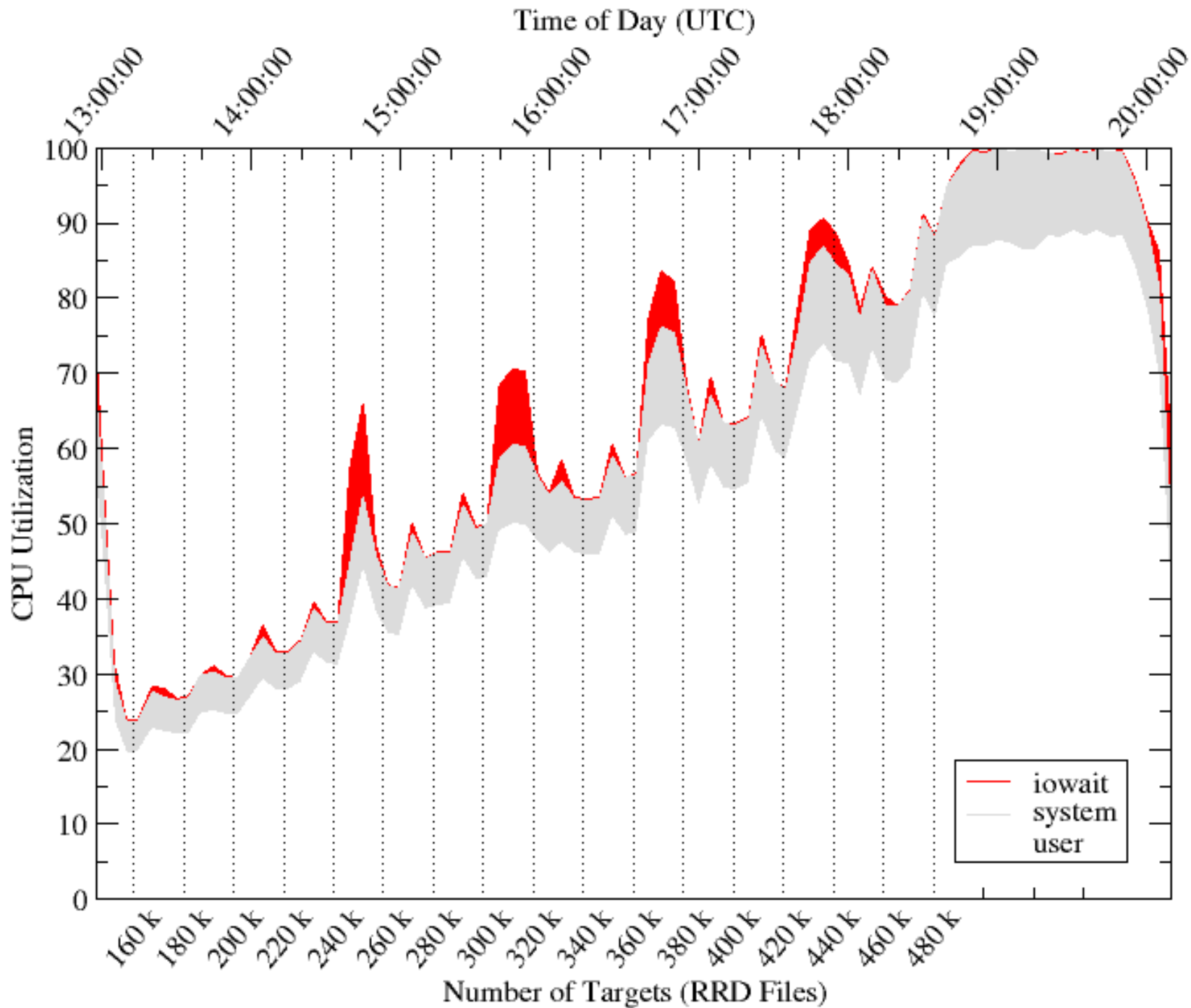
Scalability: `fadvice` (RANDOM)



Scalability: RRDCache + **fadvise**



Scalability: RRDCache + **fadvise**



Contributions

- A **method and user tools** to examine buffer-cache and readahead behaviors:
 - **fincore**
 - **fadvise**
- An **analytical model** and simulation of page fault behavior for RRD files:
 - Useful to size memory and to determine system capacity
- **RRDTool performance optimizations:**
 - Application-level buffering: RRDCache
 - Application advice: **fadvise(RANDOM)**
 - Result: approx. triples system capacity

Thanks!

- Thanks to:
 - Hideko Mills
 - Robert Plankers
 - Kevin Kettner II
 - Michael Swift
 - Tobi Oetiker
 - Released rrdtool-1.2.24 yesterday (Nov 13, 2007) with our **fadvise(RANDOM)** patch (Patches available for 1.0.49 and 1.2.23: See “Conclusions” in paper.)
- “Questions?”