

Internet2 NetFlow Weekly Reports

Stanislav Shalunov

TDD Briefing, Ann Arbor, 2004-10-08

Methodology of Data Collection

- Collect 1% sampled NetFlow data from all core Abilene routers
- Collection done at ITEC-Ohio with `flow-tools`
- Throw away data coming from interfaces between core nodes
- `flow-tools` now include SNMP hooks for that
- Concatenate the rest of the data
- Ship the resulting files (20–25 GB) to our RAID array daily
- Resulting view treats Abilene as a single data-forwarding unit

Methodology of Data Processing

- The goal is to capture long-term trends
- Weekly averages for everything, hence weekly reports
- Daily averaging too volatile, monthly would take too long
- Two data sets: one the complete thing, one “bulk TCP”
- Bulk TCP is a TCP connection that transferred > 10 MB
- For full data set can do traffic composition
- For bulk TCP data set can do more, including throughput
- Traffic composition studies are routine (though most do not look at file sharing), but looking at bulk TCP is unique

Data Presentation

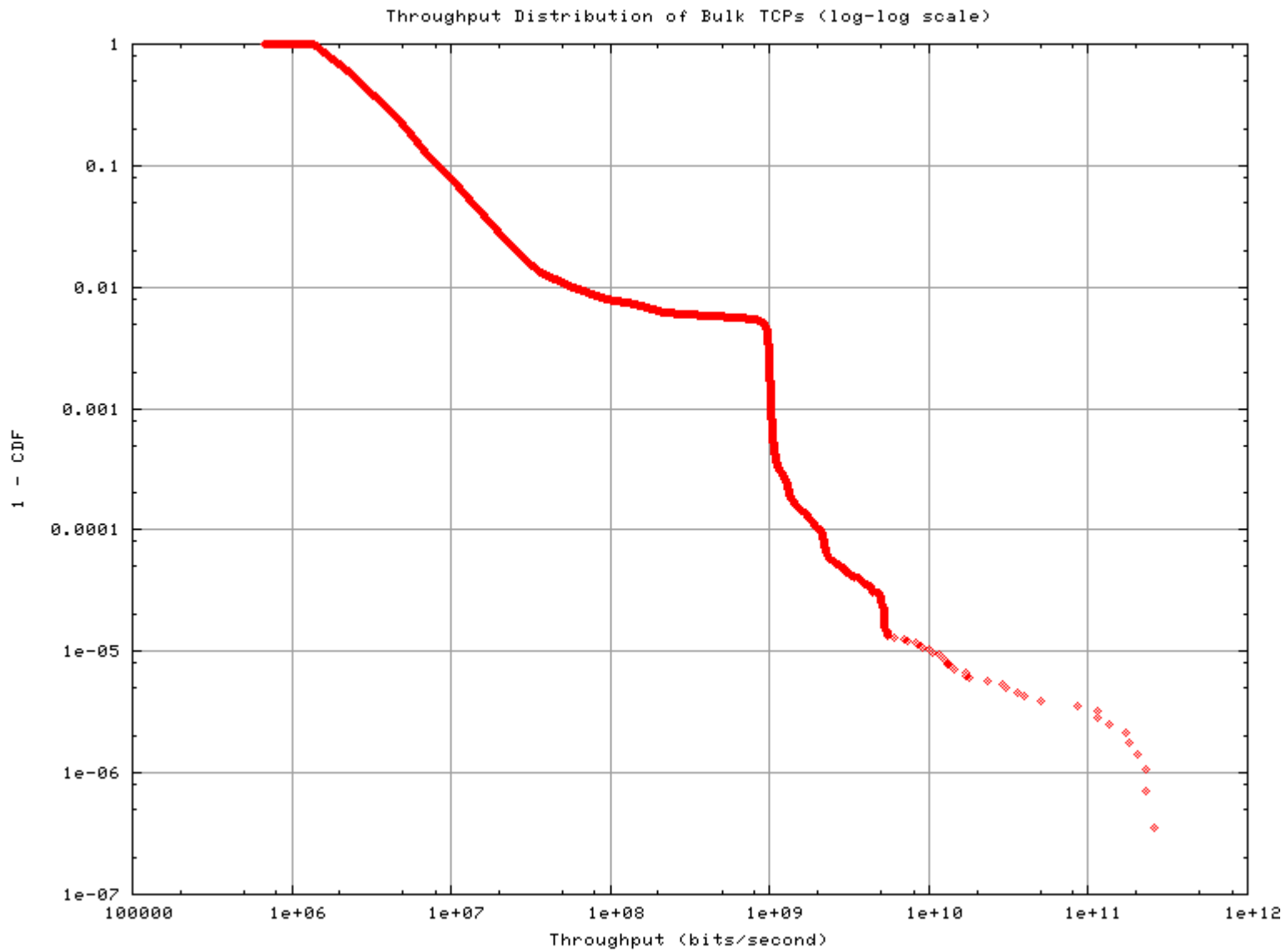
- Find it at <http://netflow.internet2.edu/weekly/>
- Weekly: new report added, time-series graphs updated
- The heart: TCP throughput analysis (includes CDF)
- Time-series graphs
- Traffic composition
- Salient points:
 - Median bulk TCP throughput is around 3 Mb/s
 - 95th percentile is around 10–15 Mb/s
 - 25-50% of traffic is measurement
 - A decreasing fraction of traffic is file sharing
 - Bulk TCP throughput appears to be increasing

Top 10 Connections

- Top 10 bulk TCP performers
- Only a single connection from a given AS to a given AS can be listed
- Two independent table are produced: one for measurement flows and one for the rest
- If going for records, check if flows show up
 - Validation
 - Independent verification

Costs, Tools Used

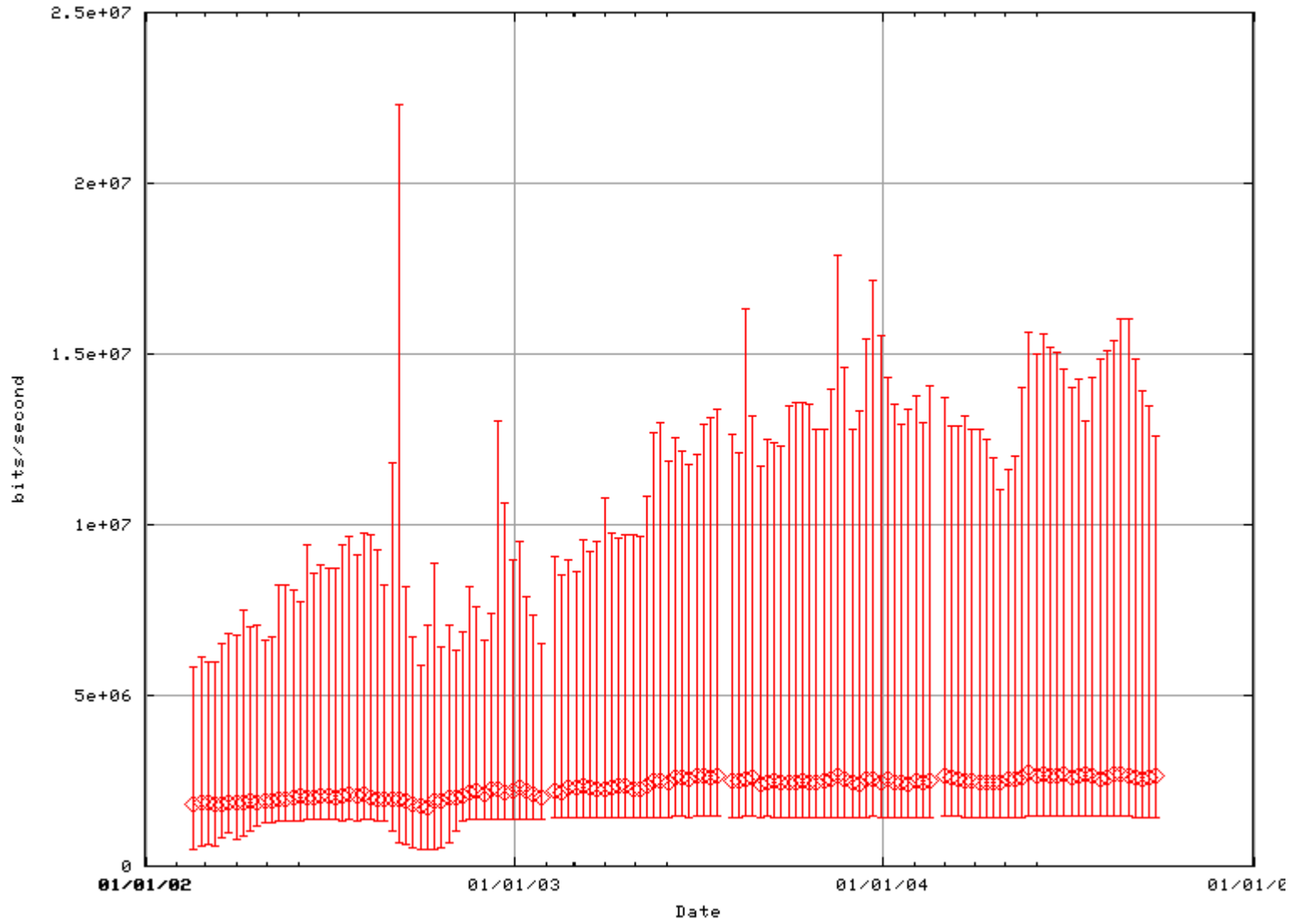
- Capacity overhead of data collection is negligible
- Need a machine with disk space
- FOTS (free off the shelf) `flow-tools` for collection
- Custom-written stuff for analysis (around 2 man-months)
 - CWEB program to make a pass over complete data set
 - Perl programs to post-process and handle presentation
- Software available at <http://www.internet2.edu/~shalunov/nfstat/>



Throughput Distribution of Bulk TCPs

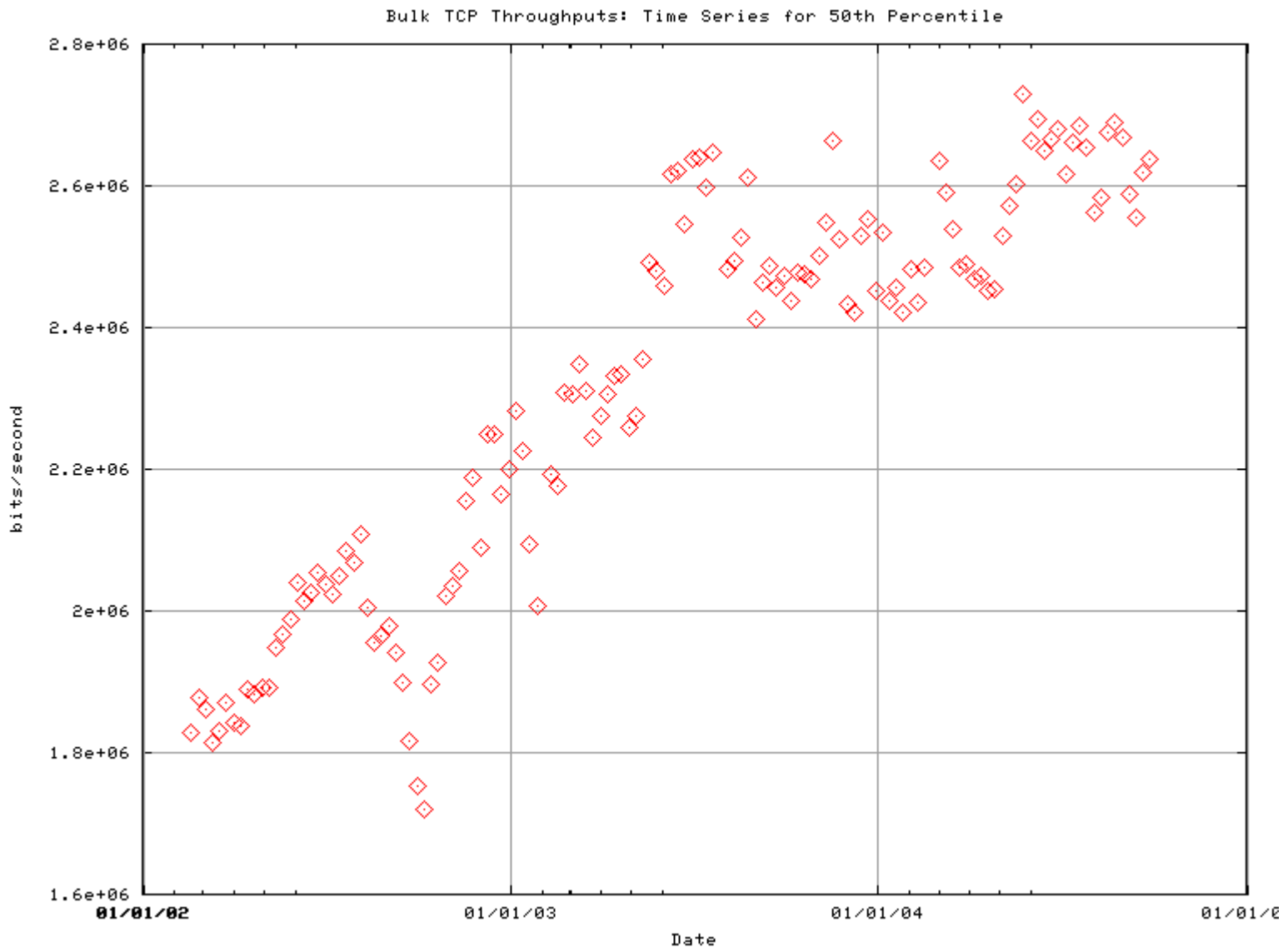
- The shape of the curve mostly similar for different weeks.
- Sometimes one encounters unusual shapes.
 - Denial of service attacks
 - Sets of major demos
- The tail wiggles more than the body
- The virtually straight (on log-log scale) line from 0 to roughly 100 Mb/s virtually every week
 - No explanation
 - No “theoretic” shape for this curve, but power laws are statistically common

Bulk TCP Throughputs: 5th Percentile, Median, and 95th Percentile



Median, 5th, and 95th Percentile of Bulk TCP Throughput

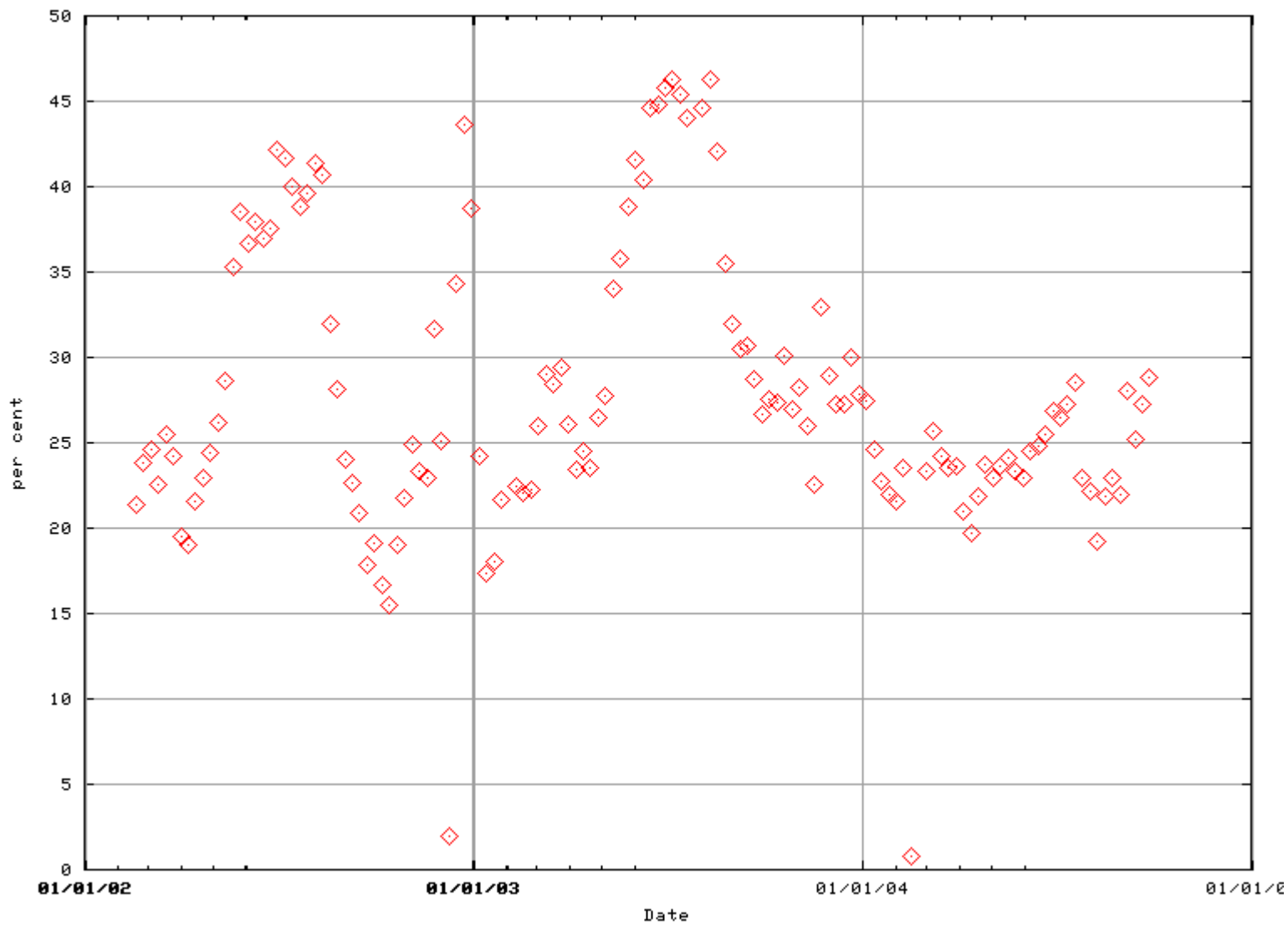
- Good news: keeps increasing (generally)
- Higher percentiles go up more than median
 - The high-end users had better luck than the masses
 - The wizard gap is widening
- But:
 - Impact of Cisco to Juniper change (different eviction timeouts)
 - Mostly can be explained by changes in OS composition (newer OSes have larger window size)
 - Impact of file-sharing decreases (file-sharing was always below median, so decreases of file-sharing translate to increases of the median)



Median of Bulk TCP Throughput

- Same as on previous plot, but less compressed
- Clearly the trend is up

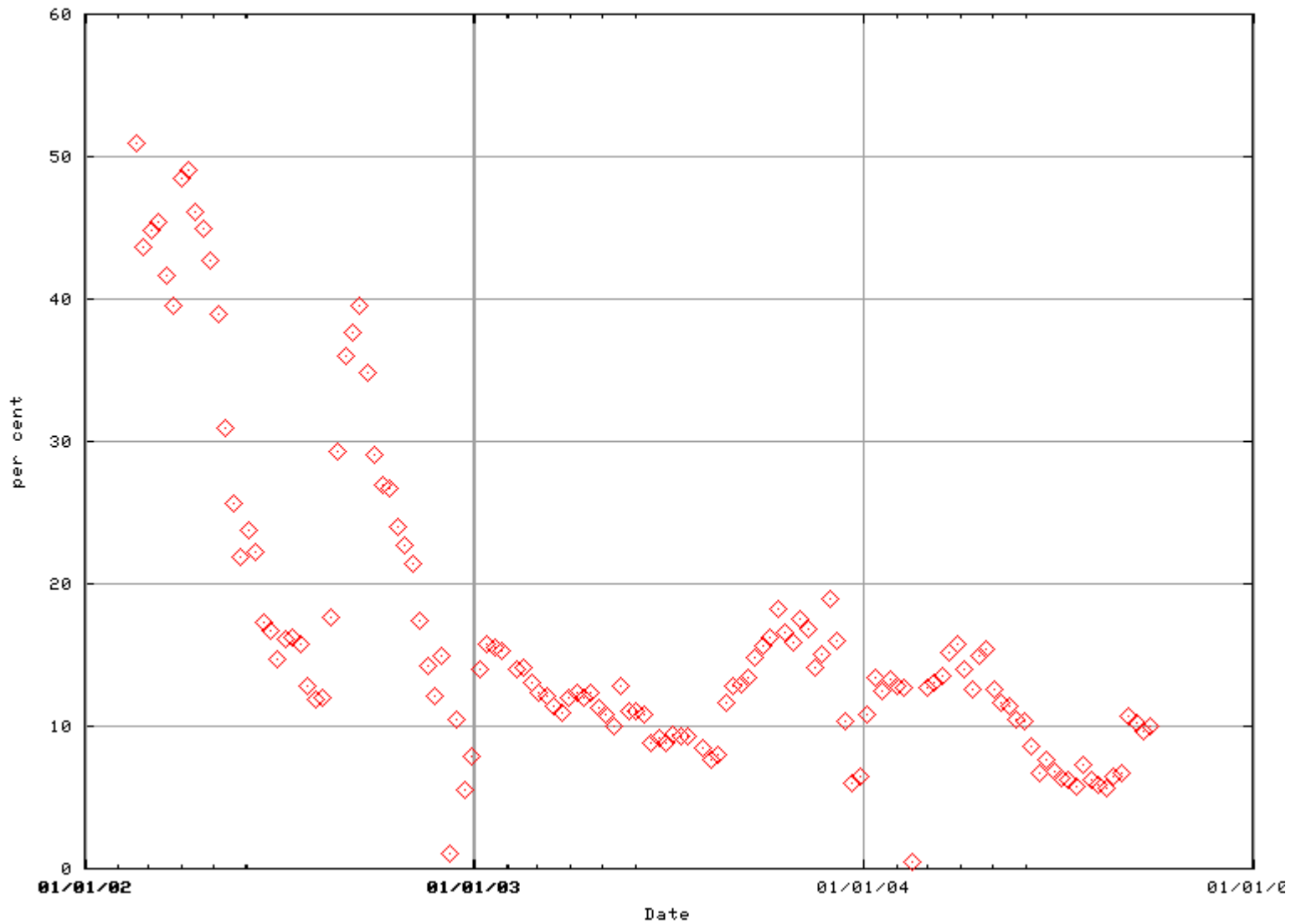
Time Series for Percentage of Data Transfers octets (Full Data Set)



Percentage of Data Transfer Traffic

- NNTP, HTTP, FTP, and Rsync
- Passive FTP not included (difficult to characterize)
- Goes up in summer and during winter break (because there's less file sharing and interactive use)
- Fairly stable otherwise

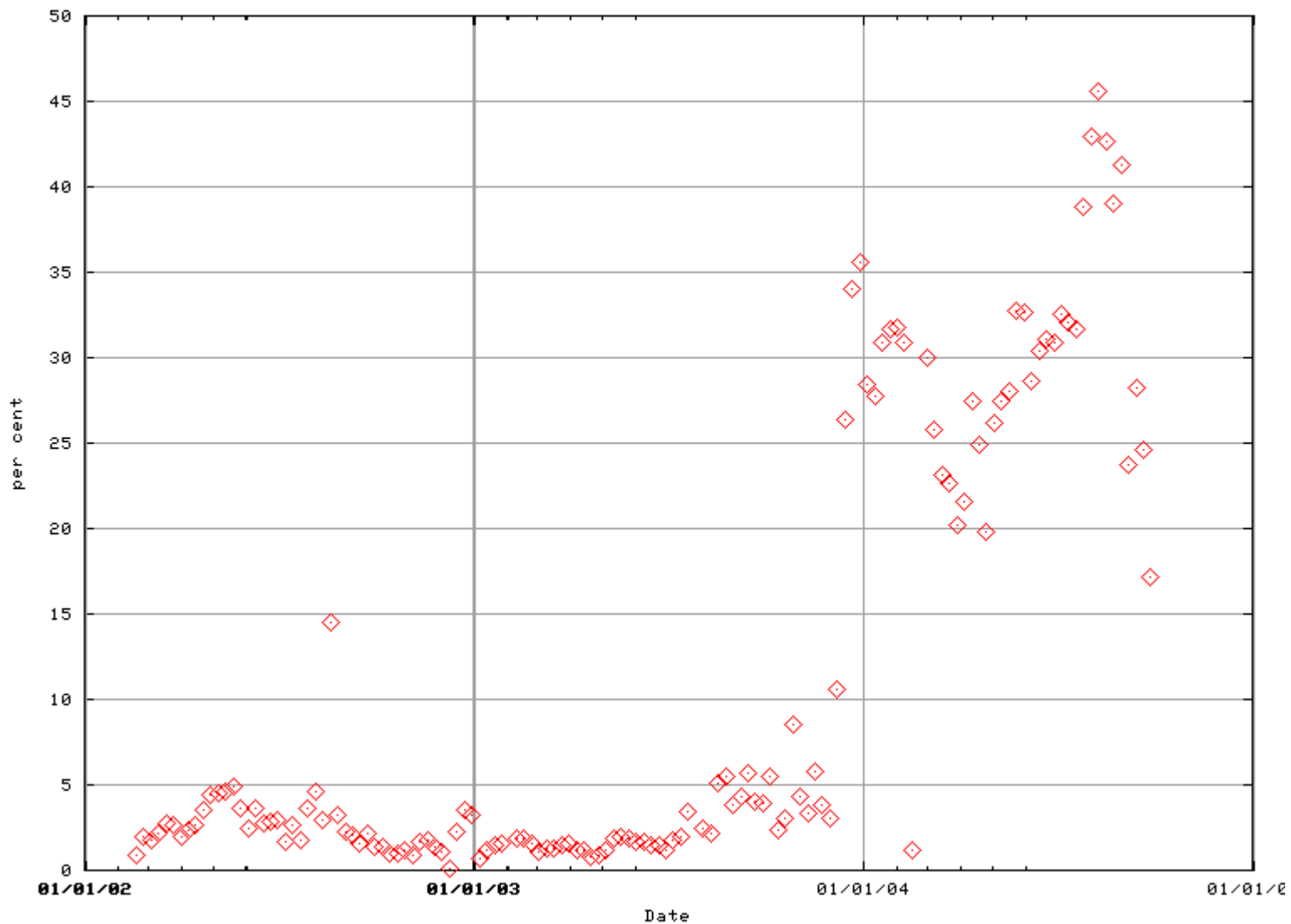
Time Series for Percentage of File Sharing octets (Full Data Set)



Percentage of File-Sharing Traffic

- The general trend is downwards
- A lot of shifting to new applications
- Some must have spilled over into unidentified
- Correlates well with political and legislative events

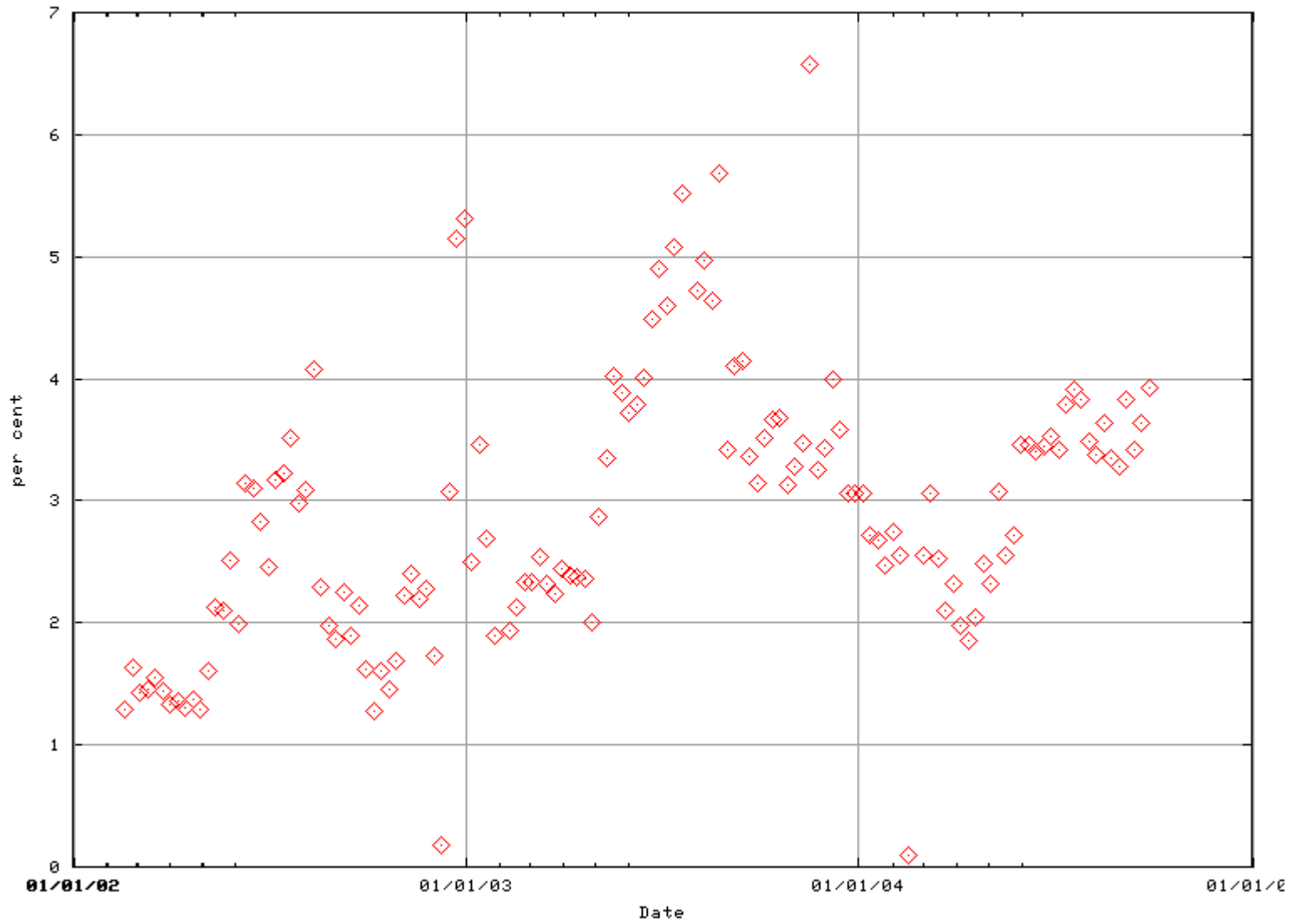
Time Series for Percentage of Measurement octets (Full Data Set)



Percentage of Measurement Traffic

- Iperf, ICMP, IPMP
- A sizeable chunk of network capacity
- Probably more than desired

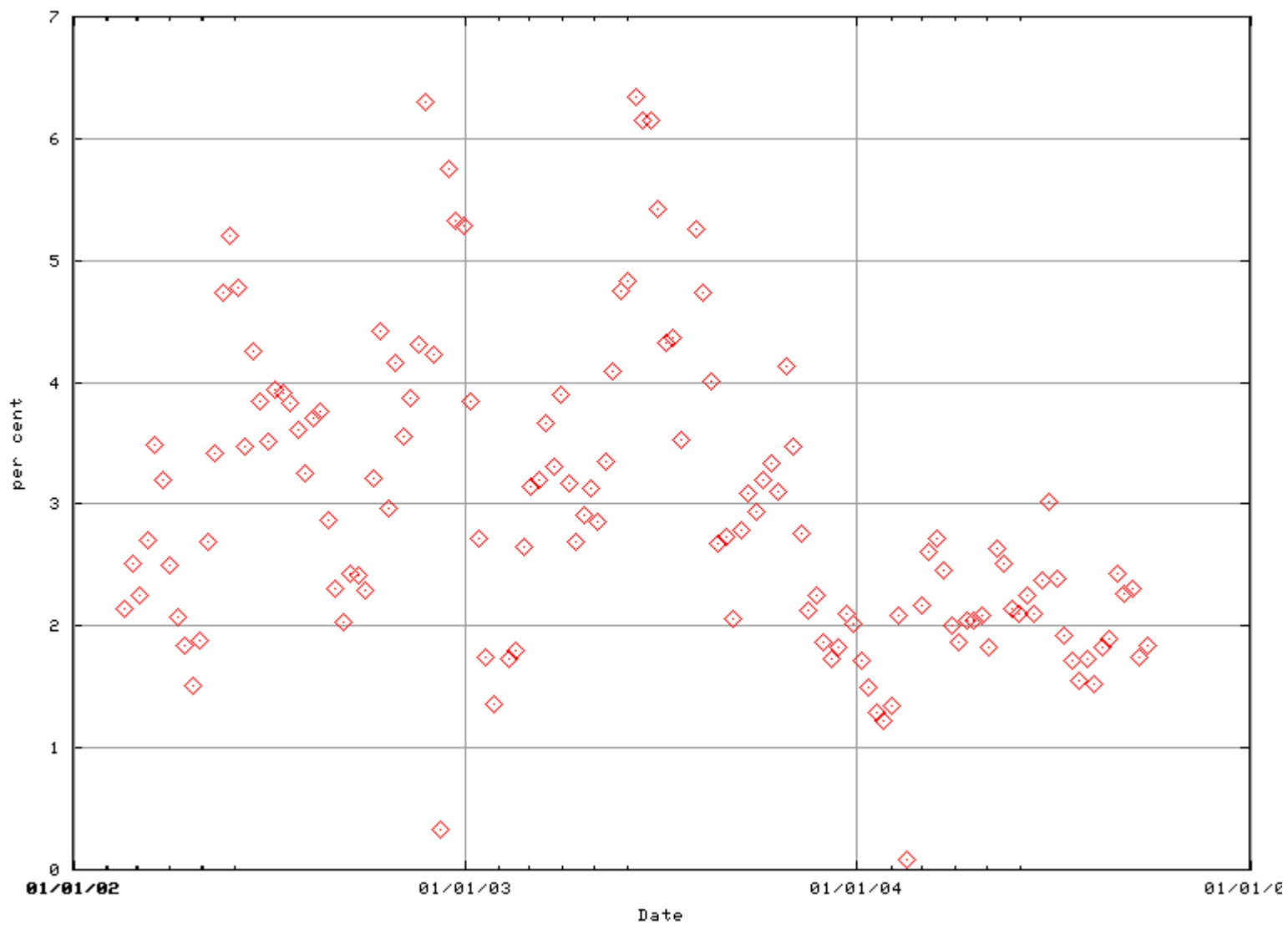
Time Series for Percentage of Encrypted Traffic octets (Full Data Set)



Percentage of Encrypted Traffic

- SSH, HTTPS, IPsec
- General trends seems to be up
- Looking at port numbers, so don't know what fraction of unidentified is encrypted

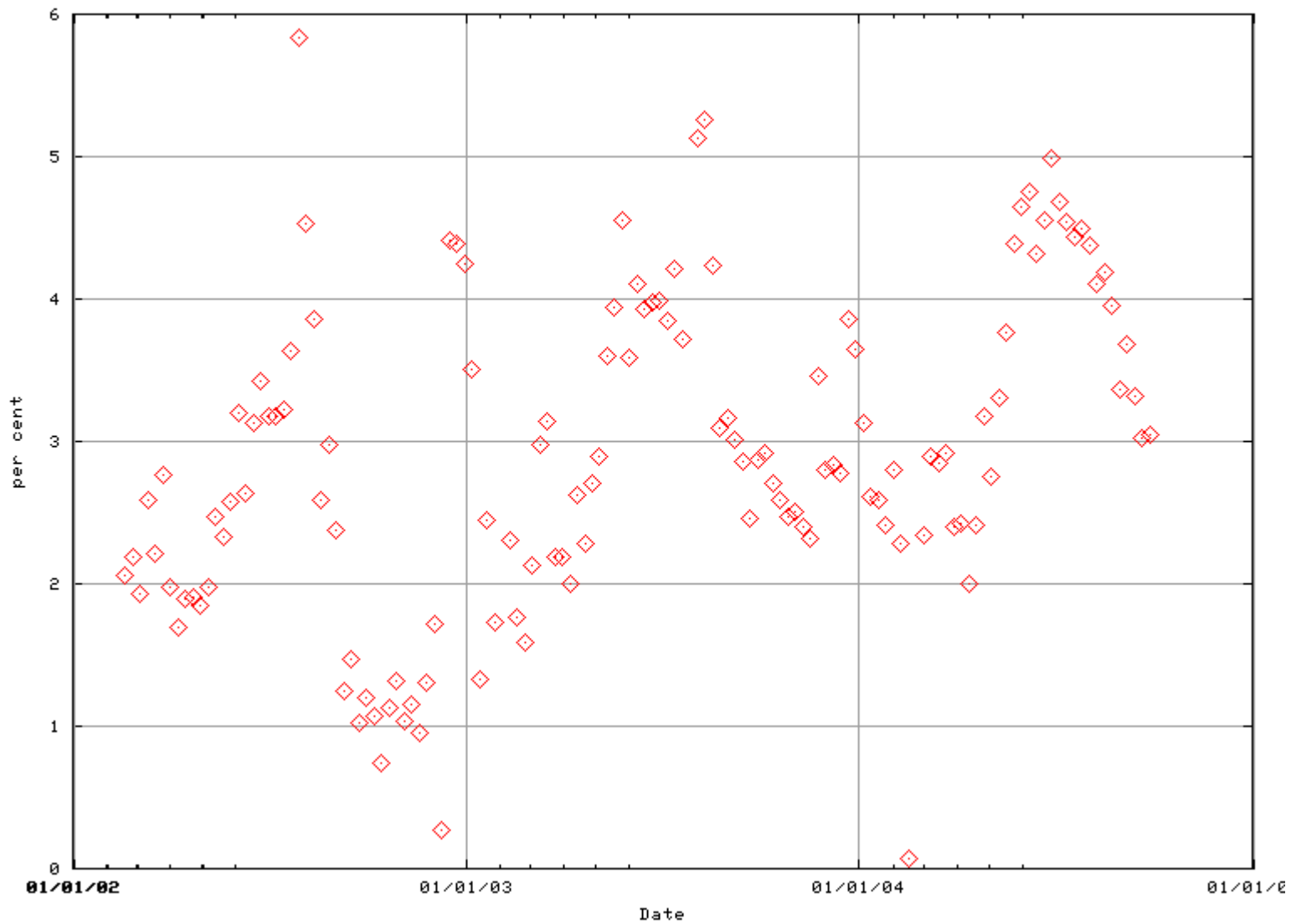
Time Series for Percentage of Audio/Video octets (Full Data Set)



Percentage of Audio/Video Traffic

- Multicast, Real, Windows Media, etc.
- More volatile than other traffic categories
- Event-related
- Windows Media a minor (but increasing) fraction of Real

Time Series for Percentage of Advanced Apps octets (Full Data Set)



Percentage of Advanced Applications Traffic

- UNIDATA LDM, BBFTP, IBP, GsiFTP, McIDAS
- Mostly LDM with a smidgen of BBFTP
- Even more volatile than audio/video
- The fewer users, the more volatility, generally

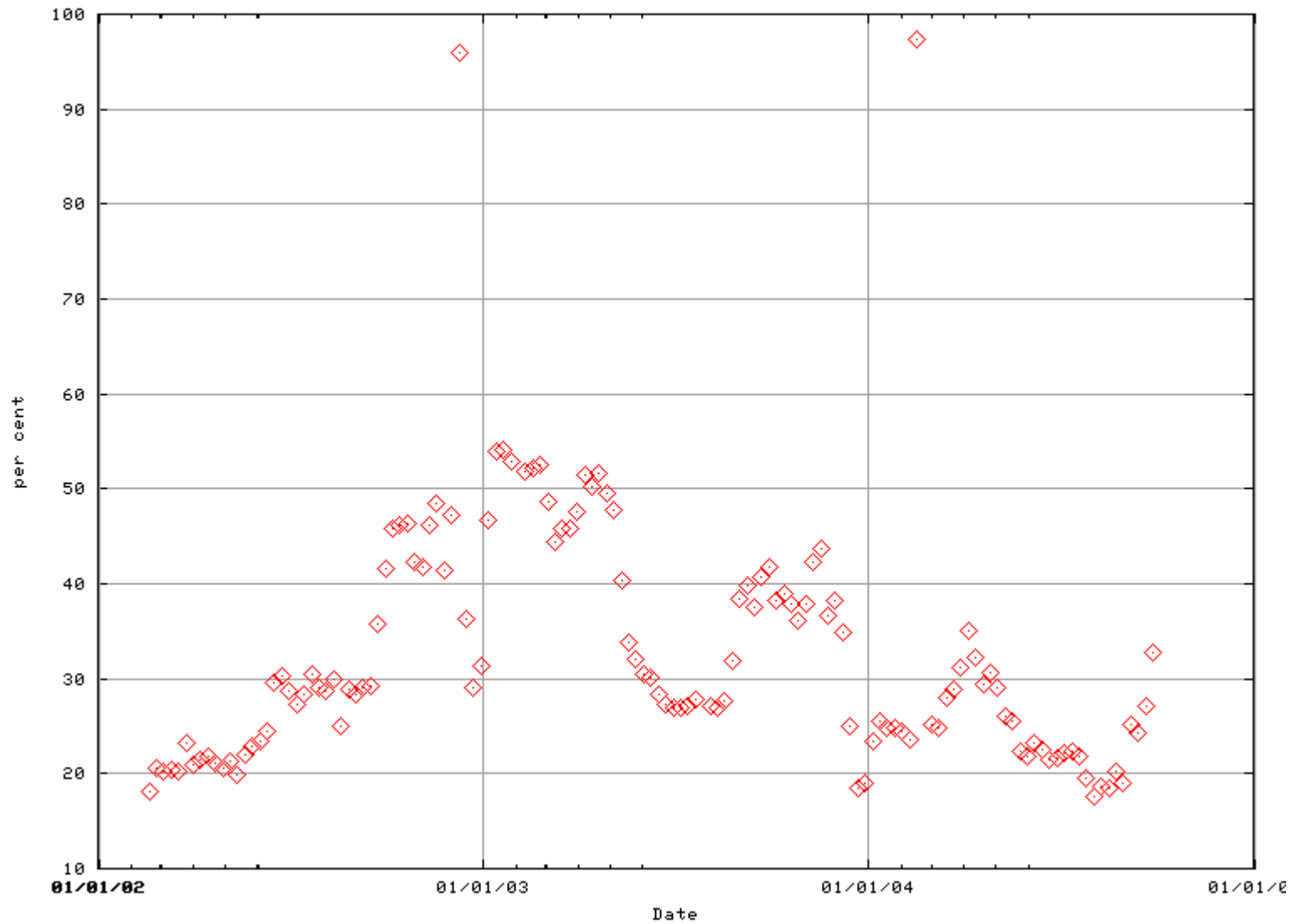
Percentage of Games Traffic

- Not a big traffic source
- Games are generally designed for the masses
- The masses have DSL or cable at best
- Most games are designed so that the bandwidth of dialup is almost enough and DSL or cable are enough
- Advanced application waiting to happen?

Percentage of Miscellaneous Traffic

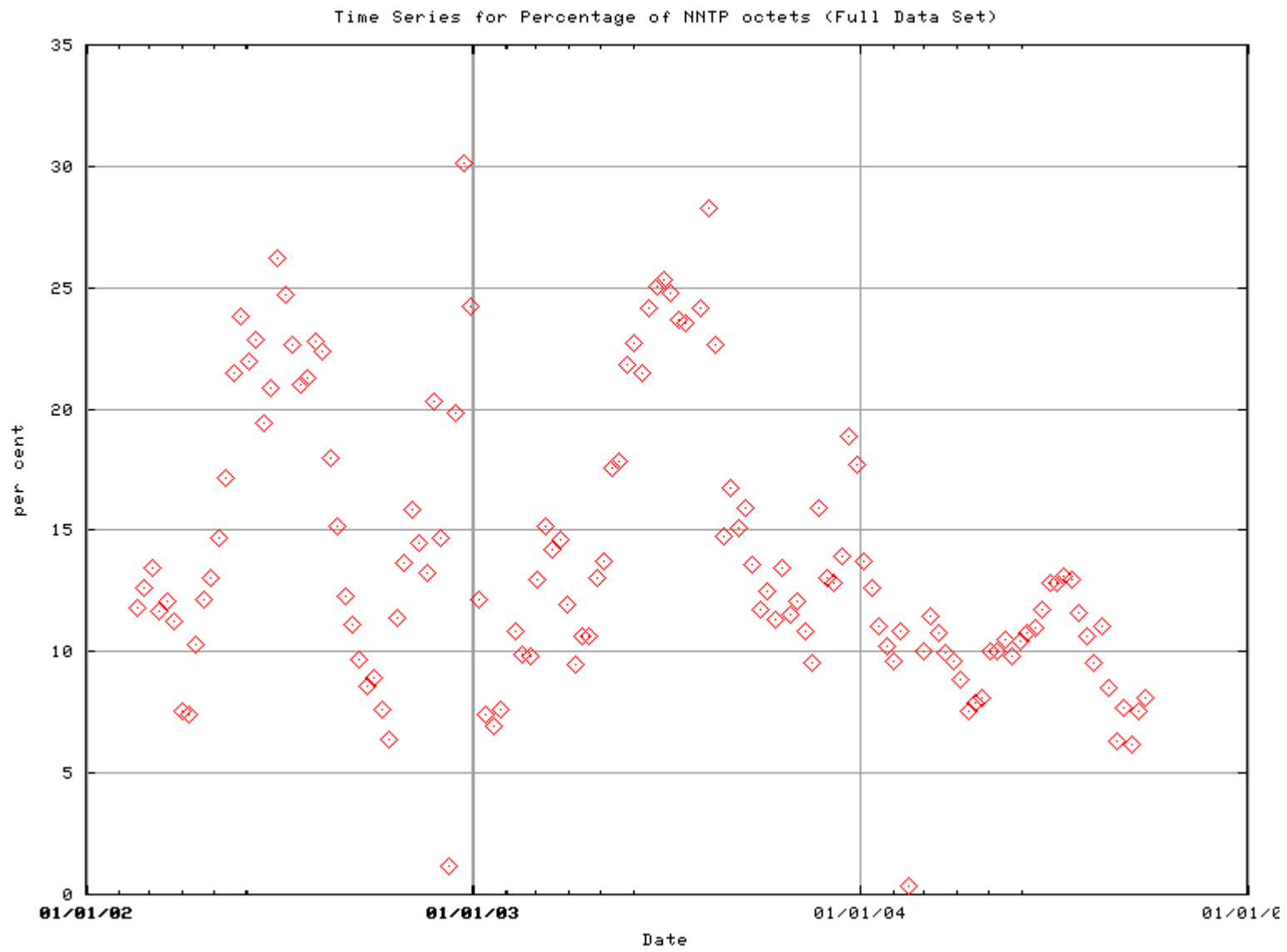
- Mail, Port 0, AFS, DNS, X11, AIM, Telnet, MS Windows, Squid, NFS, SOCKS, IRC, IDENT, NTP, SNMP, Portmapper, RTIP
- Known traffic that doesn't fit other categories
- Quite stable

Time Series for Percentage of Unidentified octets (Full Data Set)



Percentage of Unidentified Traffic

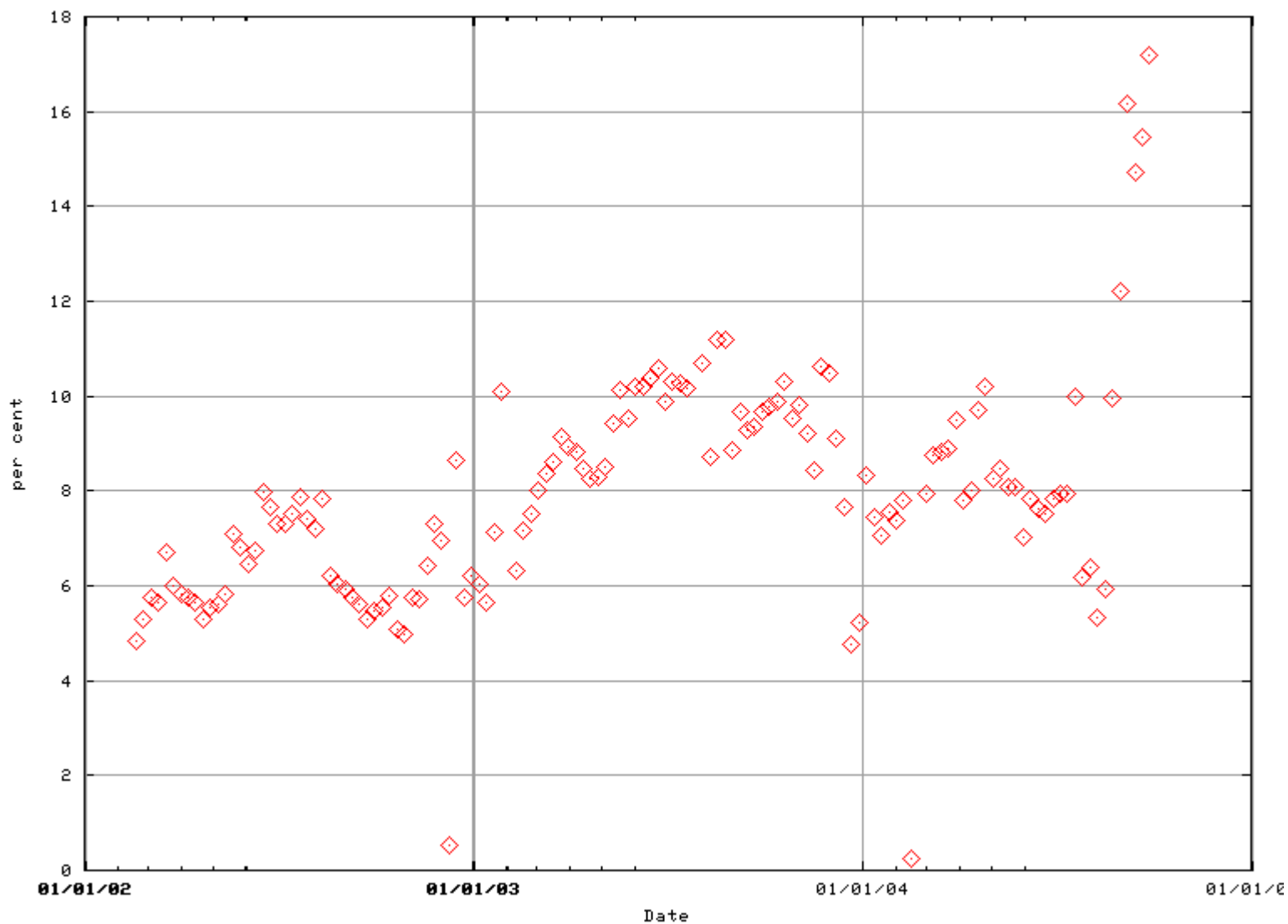
- Quite smoothly changing
- Seems to be negatively correlated with file-sharing in winter
- Seems to be positively correlated with file-sharing in summer
- An unknown fraction might be file-sharing



Percentage of NNTP Traffic

- Used to be the most talkative single application on Internet2
- Now behind Iperf and HTTP
- Some percentage swings are related to file-sharing changes
- Most bytes are in binary groups

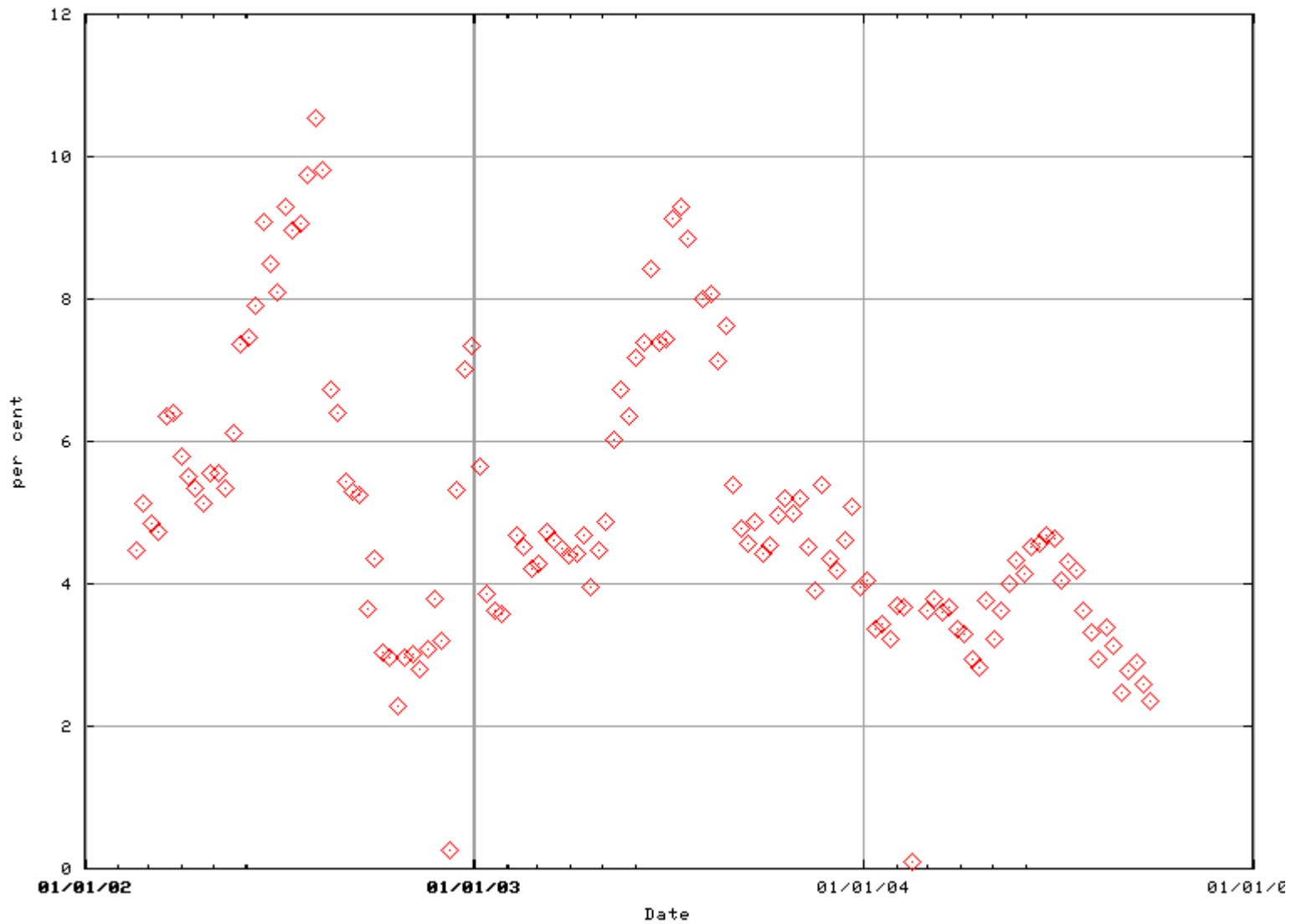
Time Series for Percentage of HTTP octets (Full Data Set)



Percentage of HTTP Traffic

- Less on Internet2 than on commodity Internet
- Not a lot of traffic (in relative terms), huge utility
 - Email has even less traffic and even more utility
- The upward trend is mostly just the decrease in other types of traffic
- Is likely to decrease, long-term (but likely retain the utility, if similar to email)
- The recent spike: a new file sharing app?

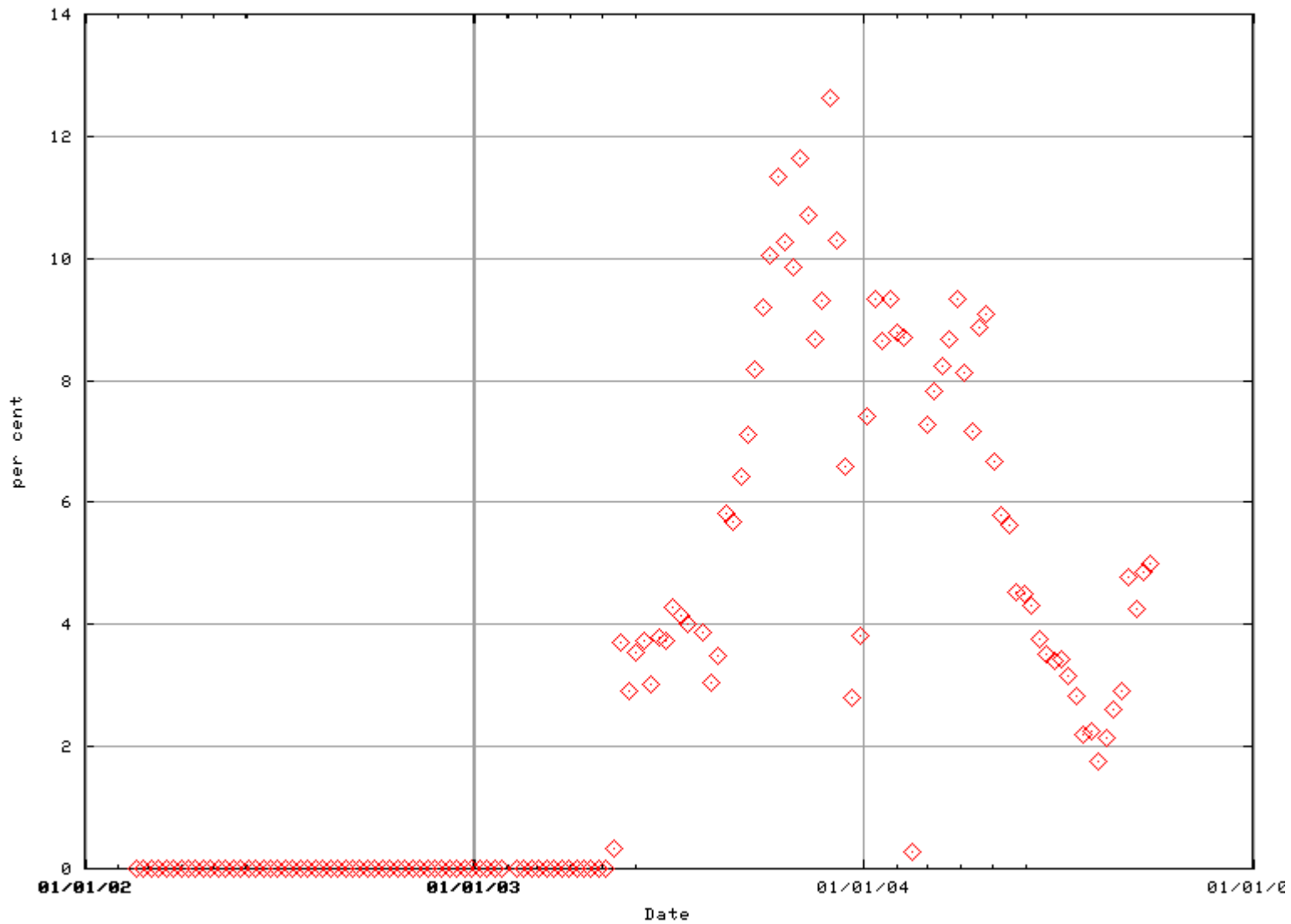
Time Series for Percentage of FTP octets (Full Data Set)



Percentage of FTP Traffic

- Active FTP only
- Passive FTP is a part of unidentified
- Peculiarly, percentage decreases during breaks
 - Therefore, mostly used interactively
- General trend seems to be down
- Seems to be negatively correlated with HTTP

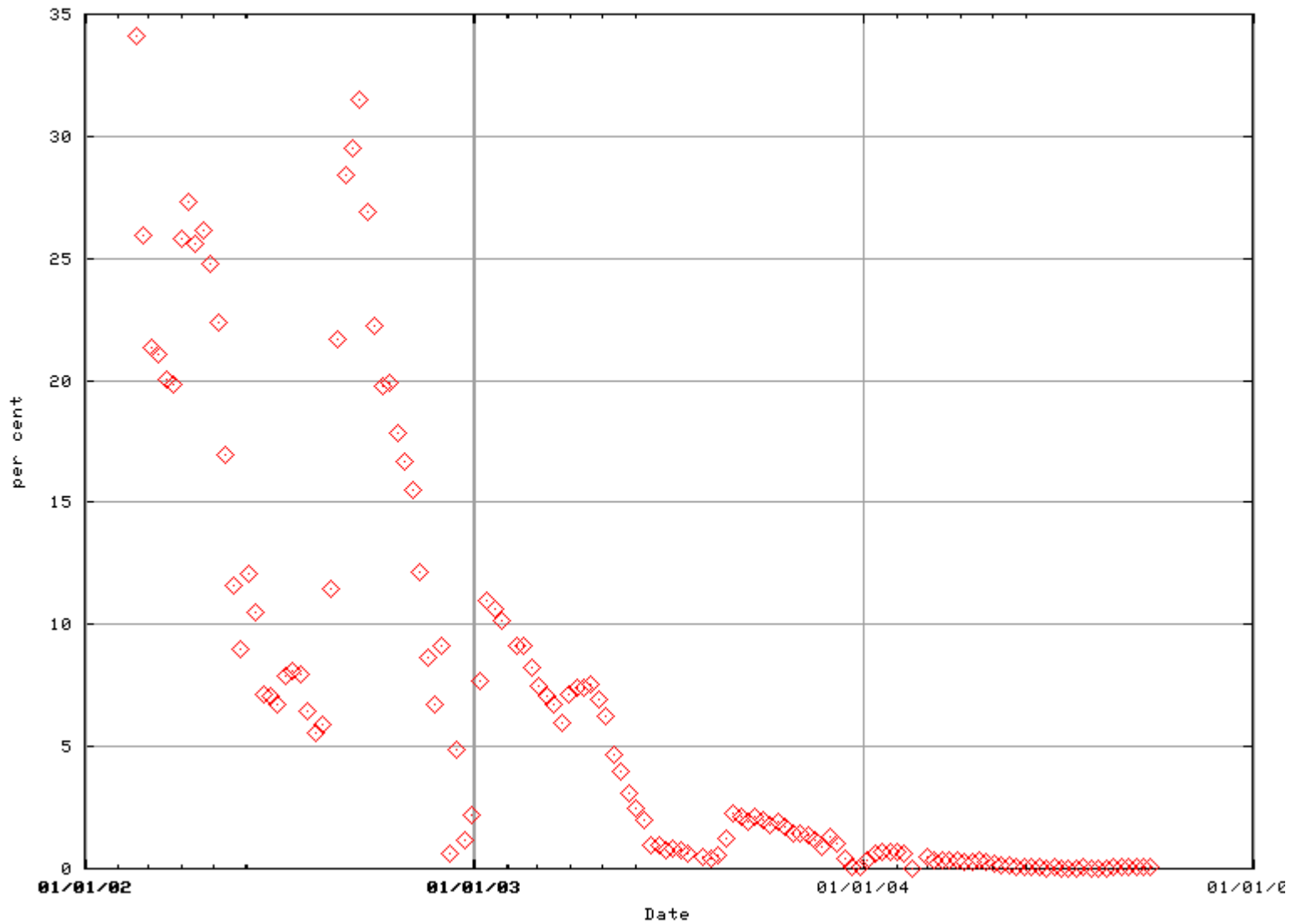
Time Series for Percentage of BitTorrent octets (Full Data Set)



Percentage of BitTorrent Traffic

- A relatively new file-sharing application
- Open-source
- Originally developed for the distribution of Linux CD images
- Used for other kinds of files nowadays

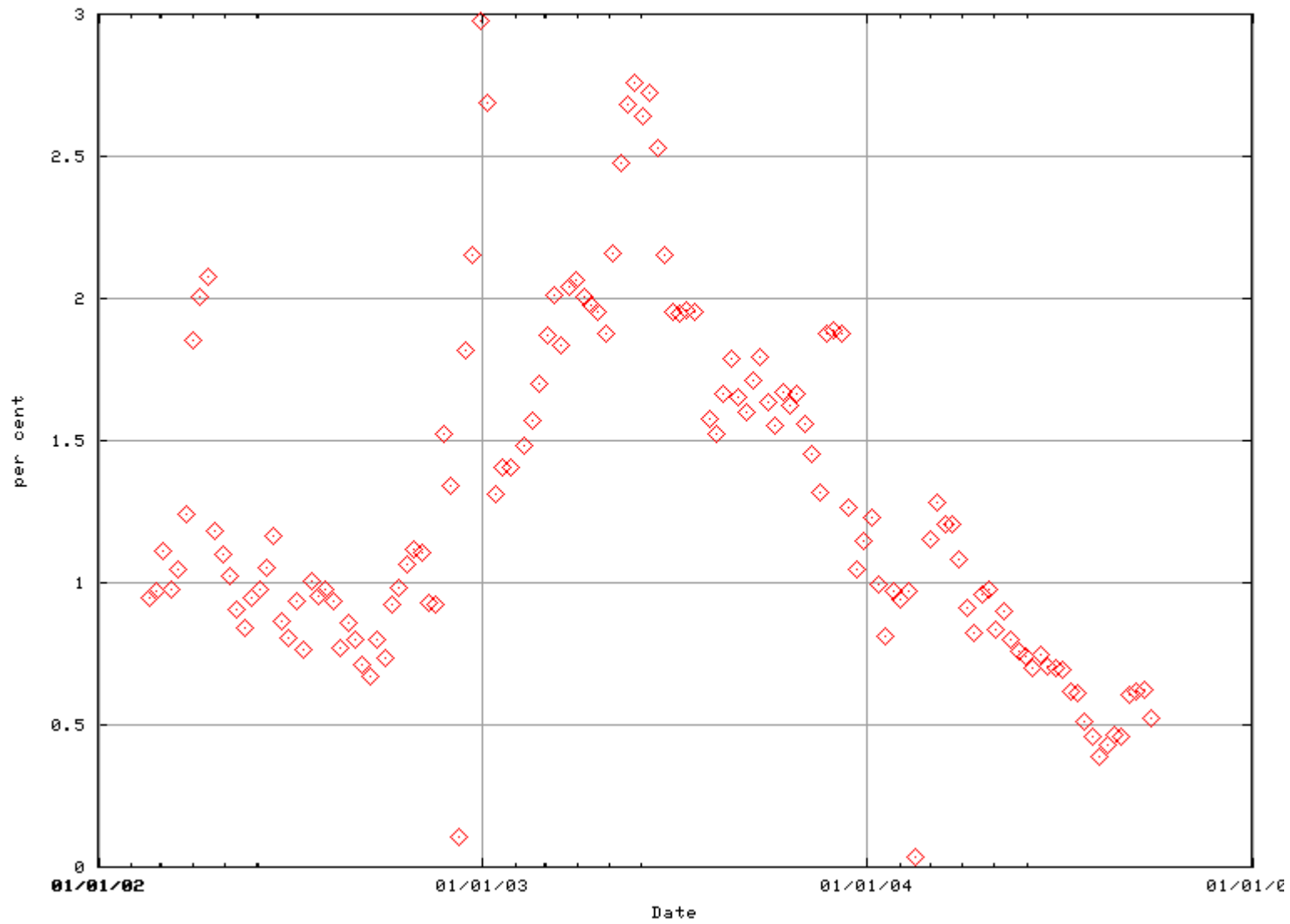
Time Series for Percentage of FastTrack octets (Full Data Set)



Percentage of FastTrack Traffic

- The original protocol used by KaZaa
- Gone and unlikely to come back in the original form

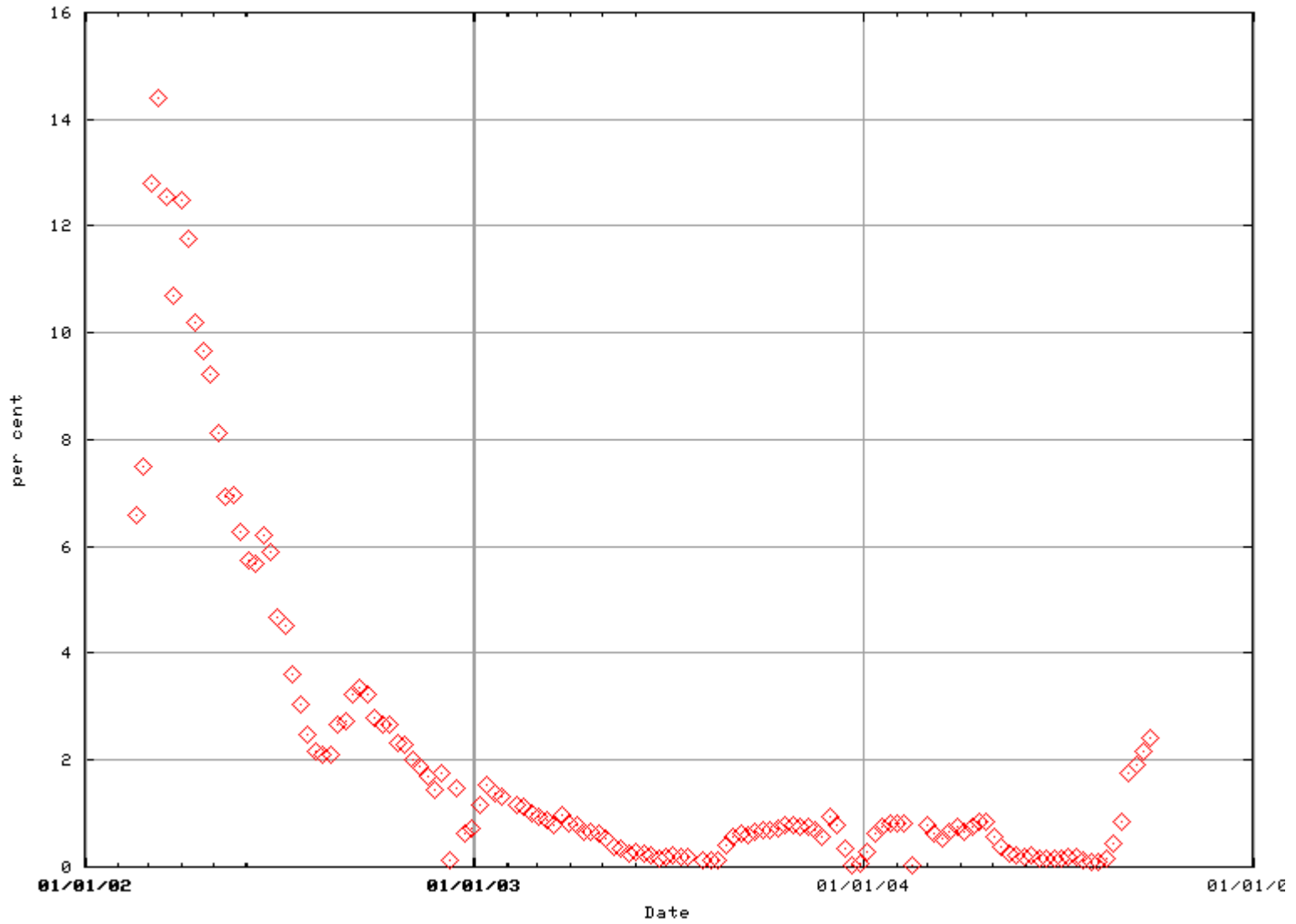
Time Series for Percentage of eDonkey2000 octets (Full Data Set)



Percentage of eDonkey2000 Traffic

- A second-tier file-sharing application
- Remarkably resilient
- Other file-sharing applications come and go, but eDonkey is still here

Time Series for Percentage of Gnutella octets (Full Data Set)



Percentage of Gnutella Octets

- A file-sharing application
- Used to be a KaZaa competitor
- Was steadily losing popularity
- Coming back?

Summary

- Performance is going up
- Wizard gap is widening
- Quantity of file-sharing is going down
- Top 10 tables can help in validation and verification of test results
- <http://netflow.internet2.edu/weekly/>