# Algoritmer og Datastrukturer 1

## DAIMI Greylisting

Gerth Stølting Brodal
Aarhus Universitet

# Greylisting

Greylisting er en teknologi anvendt på mail-serverne på DAIMI til at begrænse mængden af spam brugerne modtager

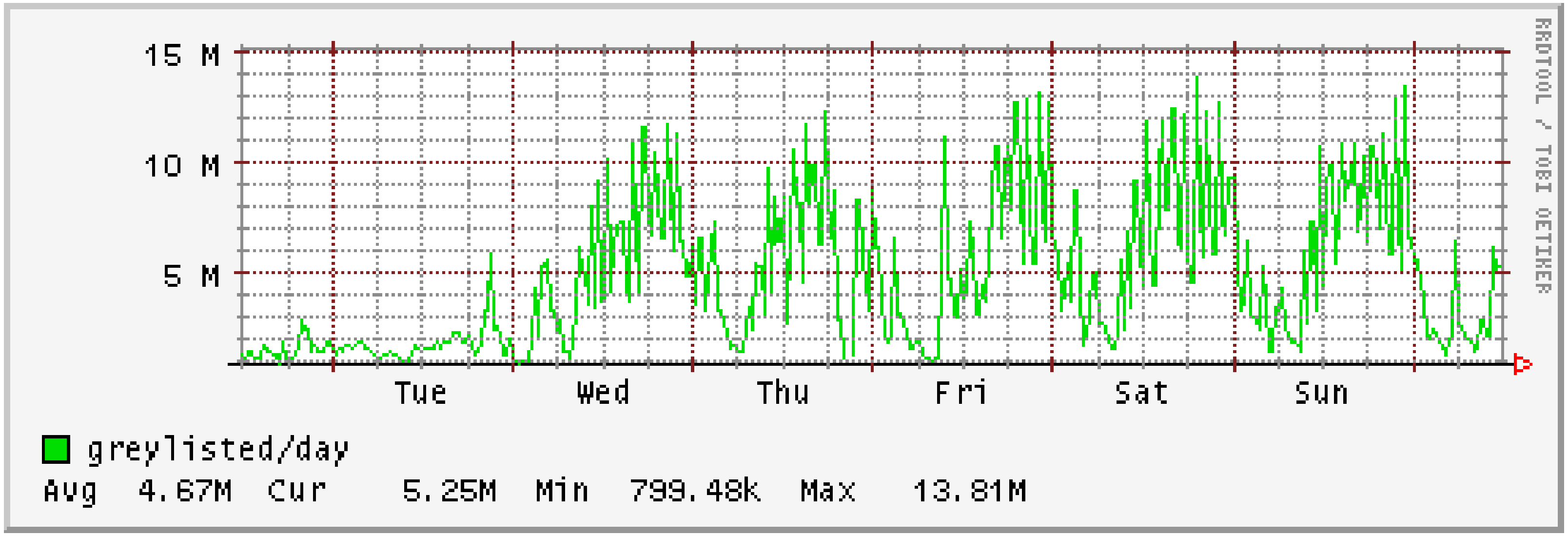
# Teknisk Forklaring

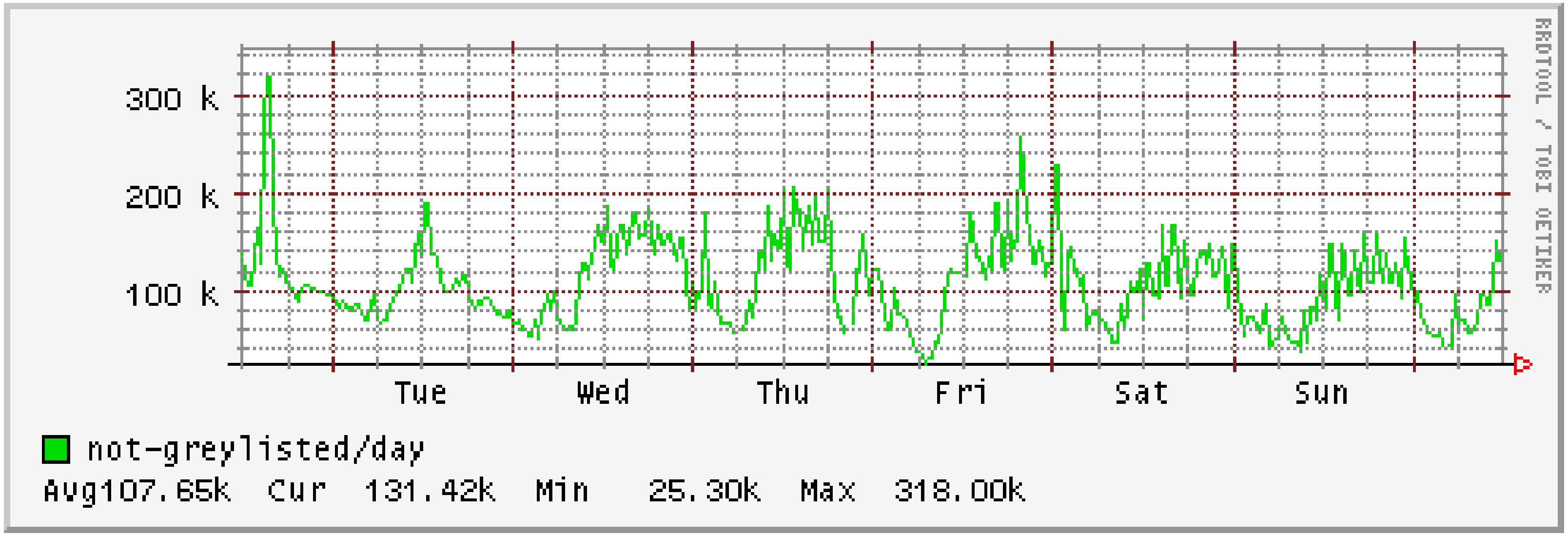
*Greylisting keeps a database of where you receive mail from. The records in the database are pentuples listing the IP network, sender domain, recipient address, a counter and a timestamp.*
*When a new pentuple is seen, the mail transaction is aborted with a temporary failure. After some time, the remote mailserver will retry the transaction. As the pentuple is in the database by now, the mail will get processed normally and reach you.*
*Mail servers (MTA's) has according to the standards to be able to handle temporary errors. However, the software used by the spammers - eg on home PCs acting as spam zombies does typically not implement this functionality.*
*Pentuples with a usage count of 1 are removed after 24 hours as they most likely represent spam.*

[http://www.daimi.au.dk/local/system/mail/grey-listing]

# DAIMI Greylisting

## February 25, 2008





# Flaskehals

- $10 \times 10^6$ mails per dag
- Gennemsnitlig 8.6 ms mellem hver email
  - kan ikke nå at skrive til disk for hver email
- Pentuple = 128 bytes
- Pentupler per dag = 1.3 Gb
  - løber tør for hukommelse

# Løsning

- Istedet for at gemme 128 bytes pentupler $p$, gem en 64 bit **hashværdi** $h(p)$
- Brug en ordbog implementeret ved **linear probing** til at gemme $h(p)$ startende søgningen på position $h(p) \bmod m$
- $10\text{x}10^6$ emails fylder $2 \text{ x } 8 \text{ x } 10\text{x}10^6$ = 160 MB

**fyldningsgraden**

$h(p)$ **fylder 8 bytes**

**# emails**

- **Pris:** Enkelte spam mails hasher til samme værdi og slipper fejlagtigt igennem greylisting