

Will P2P change storage?

Stanislav Shalunov <shalunov@bittorrent.com>

tech talk at NetApp

Sunnyvale

October 2, 2009

Talk menu

- **Aperitif: P2P design philosophy**
- **Appetizer: BitTorrent primer**
- **Main course: P2P lessons for storage**

P2P definition

- provide a service
- find peers
- choose peers
- serve peers



client/server



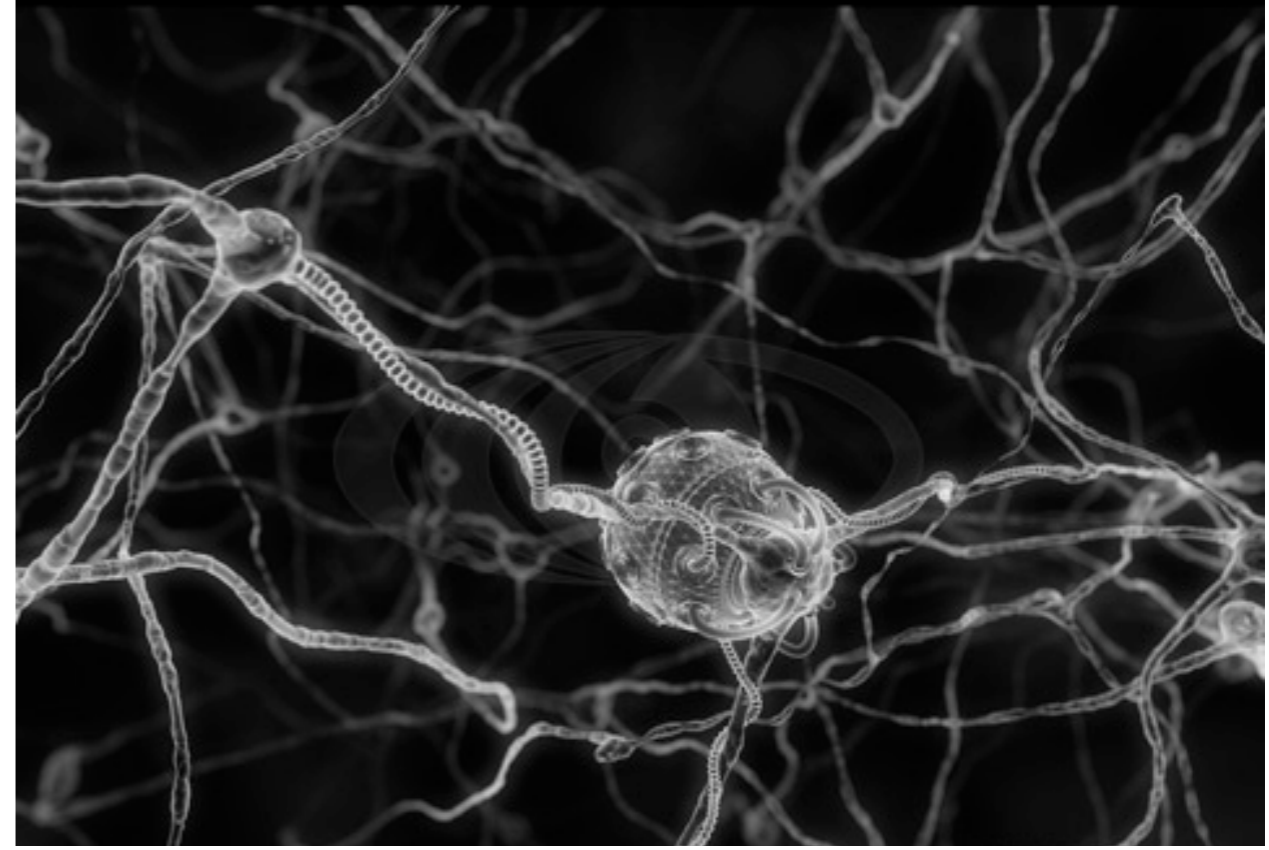
P2P

P2P properties

- self-configuring
- self-healing
- self-optimizing
- decentralize everything
- no single/double/centuple point of failure



client/server



P2P

P2P: different engineering attitude

- think living organisms, not machines
- think continuous healing, not stopping for repair
- think highly decentralized optimization, not centrally run algorithm
- think multiple redundancy, not ECC

Popular P2P apps today

- BitTorrent
- Skype

	BitTorrent	Skype
service	download	voice/video calls
protocol	open	closed
implementation	multiple	single
tracker?	yes	no
supernodes?	no	yes

BitTorrent: terms

- torrent file
- tracker
- swarm
- peer protocol
- upload slots
- rarest first
- tit for tat

Torrent file

- torrent file is to BitTorrent what URL is to the web
- metadata (file name(s), etc.)
- tracker(s)
- chop up content
- list the hashes

Tracker

- simple to implement
 - a functional tracker starts at a small PHP script
- HTTP-based
- lets you learn a few peers for a torrent
- light resource use
- not strictly speaking necessary

Swarm

- All peers sharing the same torrent
- Aim to keep well-interconnected
- Generally has a life cycle
 - Starts with too few seeds
 - Gets better and healthier on its own
 - Often goes to overseeded
 - Dies eventually

Peer protocol

- the protocol peers speak to each other
- I have a piece
- send me this
- choke/unchoke

Upload slots

- peers connect to tens, even hundreds of other peers
- most connections only transfer metadata
- connections that transfer data are controlled by upload slots
- e.g., 5 upload slots is sane

Rarest first

- Goals
 - Have something others want
 - Don't let pieces disappear
- Download rarest pieces first
- Rarity within known universe
- Known universe samples the swarm

Tit for tat

- Goals:
 - Faster downloads
 - Discourage freeloading
- Give upload slots to peers who send to us fastest
- Reserve some upload slots (e.g., 1) for acts of random kindness — optimistic unchoking

Peer exchange (PEX)

- Tell peers other peers I know
- Factor of ~100 safety margin of knowledge

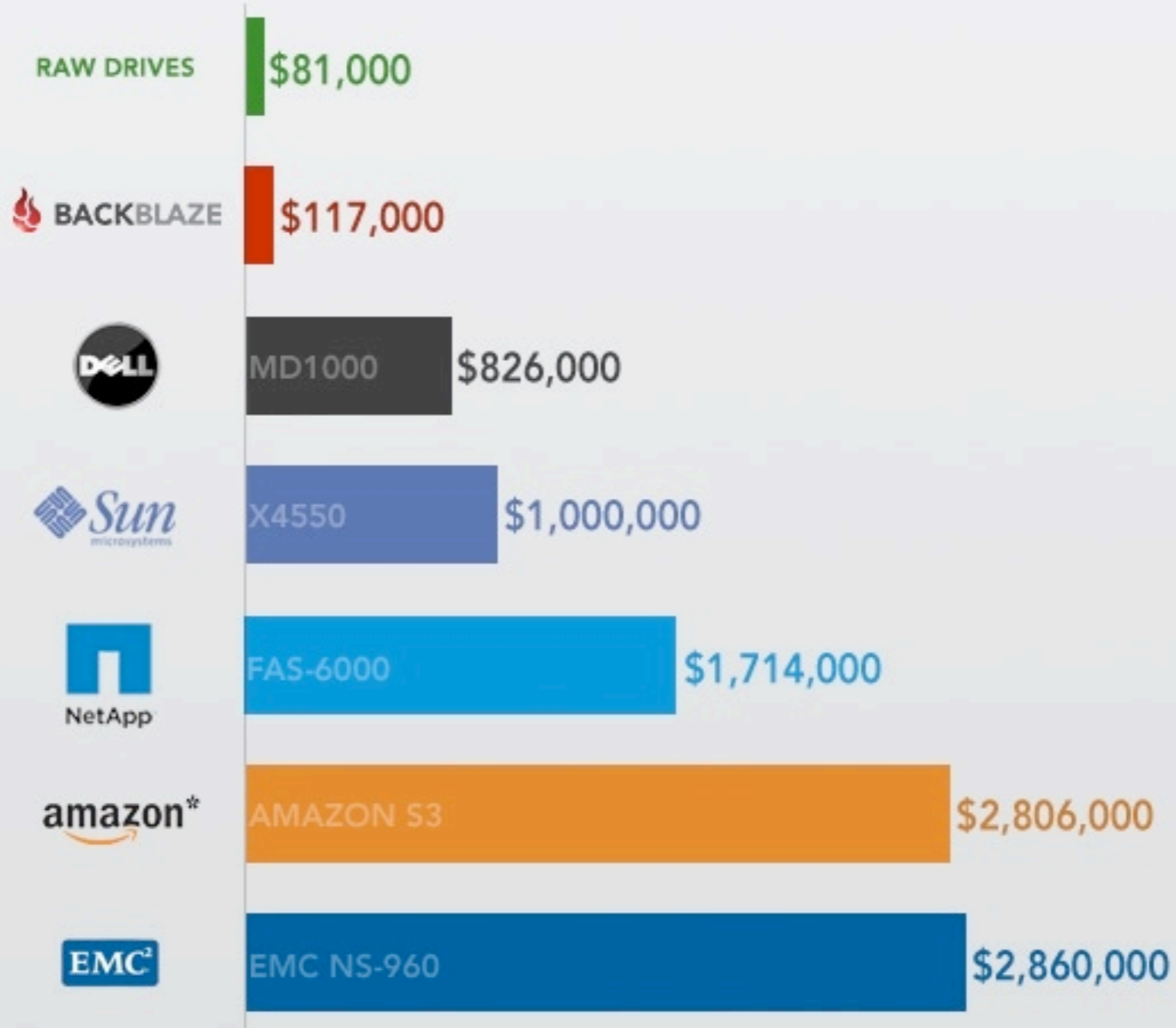
Distributed hash table (DHT)

- Makes trackers optional
- Two popular implementations

What about storage?

- Highly available storage commands a huge premium
- Raw disk capacity is cheap and getting cheaper

COST OF A PETABYTE



* Amazon S3 Storage over three years (minus electricity, co-location and administration).

comparison courtesy Backblaze

Cheaper and as reliable?

- P2P design principles produce highly reliable and highly available systems
 - No single/... point of failure
 - Healing, not repair
 - Self-optimization
- No single part needs to be reliable/available, it's a system property

Raw disk cost multiple

- commodity hardware storage: 1.3
- nice **empty spot: 3-4**
- low-end commercial storage: 10
- high-end commercial storage: 20-35

Hosted storage clouds, a red herring

- S3 is the market leader, best execution
- Very expensive (35x disk, says Backblaze)
- Not very available (goes away for hours)
- Slow to access
- Hard to access (a pain to even mount)
- Best sell so far is low startup costs

Storage cloud software

- Amazon, Google, etc., run proprietary systems
- People who buy high-end storage won't build their own
- Storage clouds (non-hosted) could be cheaper and as good
- Need software for them

Wanted:

- Throw in cheap commodity storage boxes
- Self-configure: think DHT, not config files
- Self-optimize for consistent performance
- Self-heal: think rarest first, not RAID rebuild
- Replace failed hardware at leisure
- Upgrade gradually
- Grow in small increments

Possibly useful design principles

- Equal peers (eliminates chokepoints)
- Multiple simultaneous paths for anything
- Design for peers going on and off
- Rarest first
- Stick with paths that work and always look for more
- Combine normal operations with healing?

Your turn

- Ideas?
- Comments?
- Questions?

add me on Twitter: @shalunov