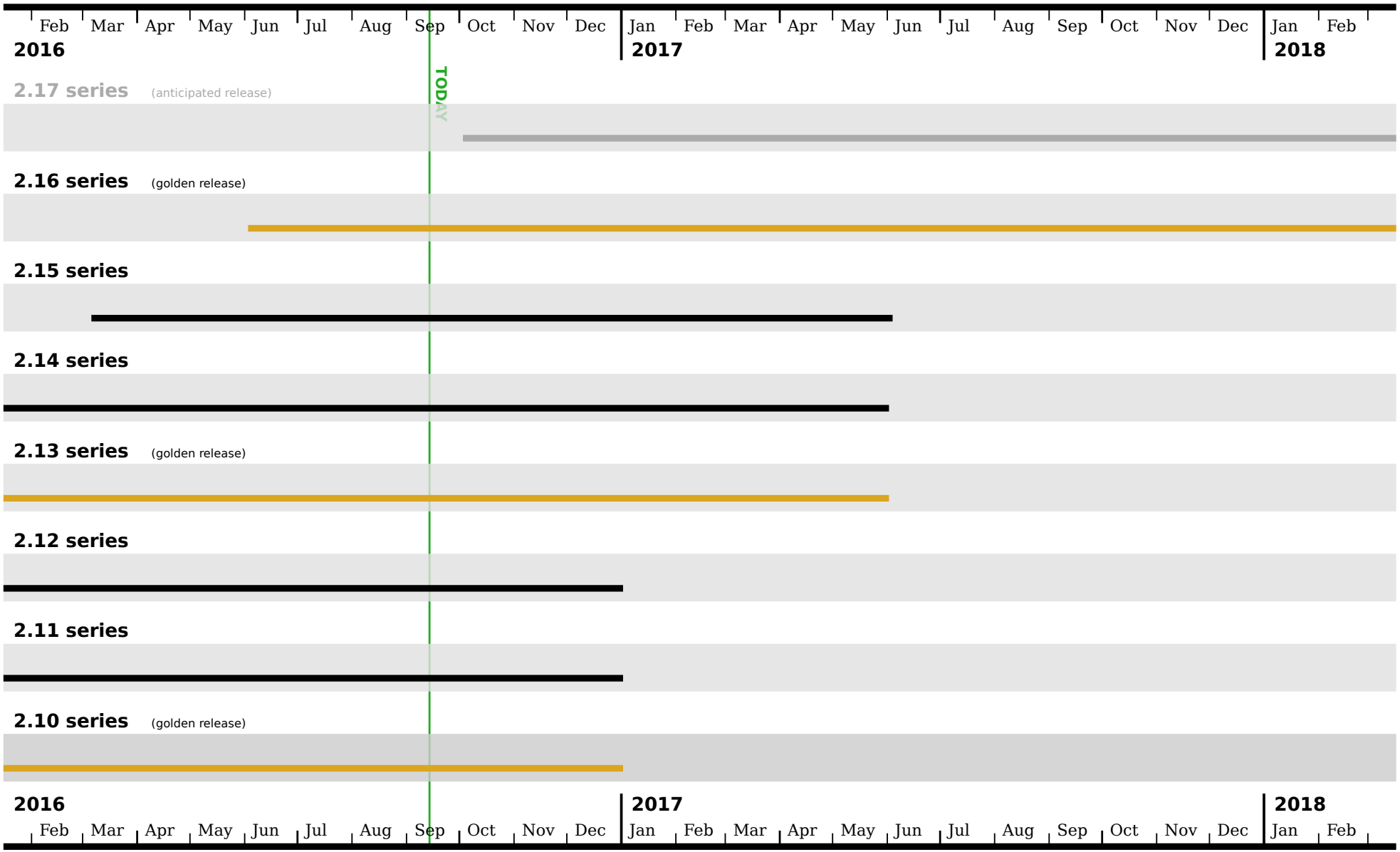
dCache

Paul Millar, on behalf of the dCache Team

pre-GDB „Data Management“ at CERN
2016-09-13
https://indico.cern.ch/event/394833/

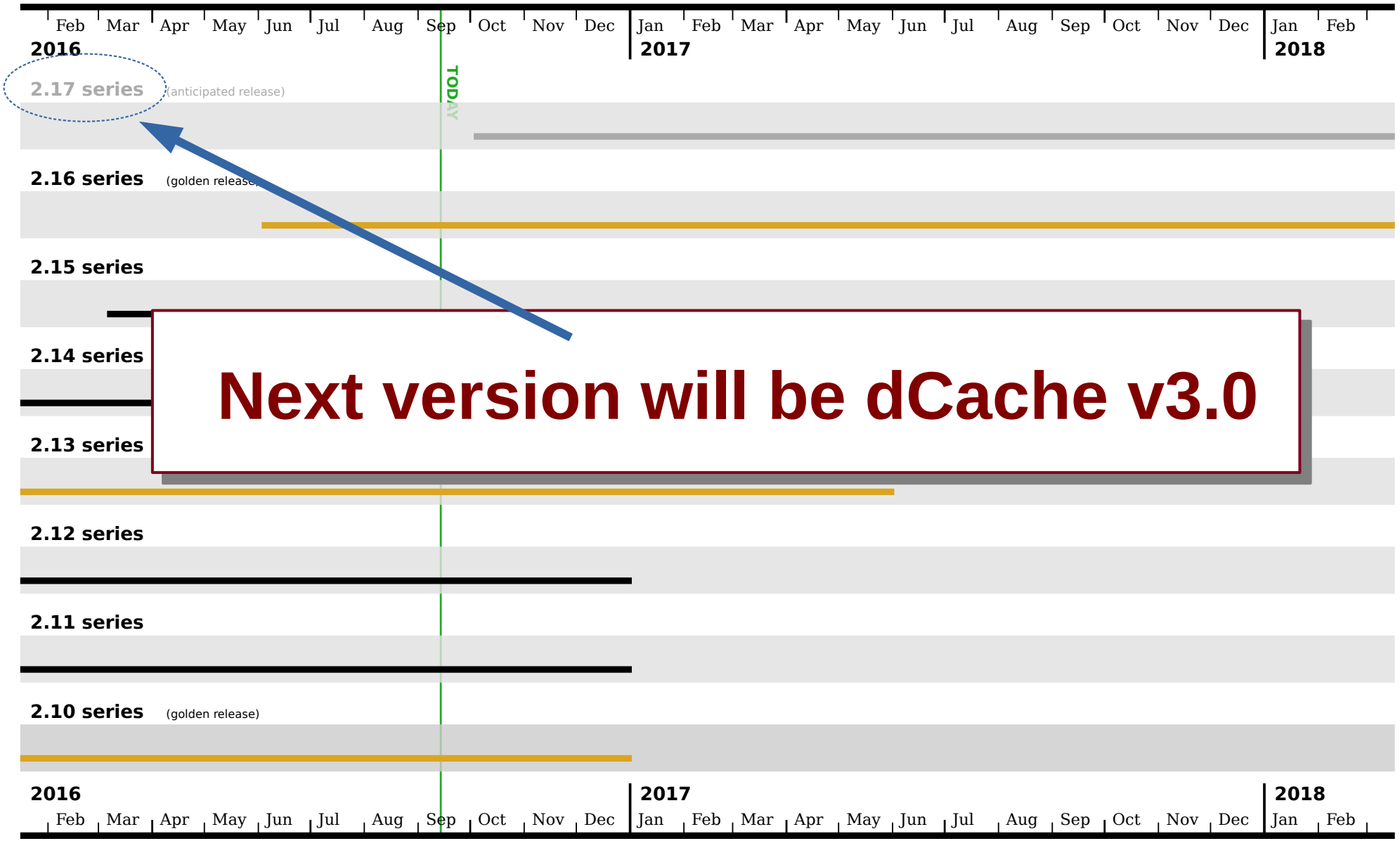
dCache server releases

... along with the series support durations.



dCache server releases

... along with the series support durations.



Why bump the version to 3.0?

- Lots of reasons (choose your favourite):
  - We have to at some point.
  - Reflects compatibility in mixed deployment.
  - Many exciting new features:

    They're optional – sites don't have to use them
- Final analysis .. just because.

New in 3.0: CEPH integration

- dCache now has built-in **CEPH integration**:

  Sites can deploy a dCache pool that provides access to a CEPH pool.

- dCache files are written as **RBD images**:

  These can also be accessed independent of dCache, if you know the PNFS-ID of the file.

- All **protocols** and **high-level features** are available:

  Sites with tape integration may need to tweak their scripts

- This is site driven functionality:

  You asked for it!

New in 3.0: HA-dCache

- **What** is HA-dCache?
  - Multiple instances of core components can run concurrently,
  - Doors updated to support load-balancers (e.g., HAProxy).
- **Why** HA-dCache?
  - Symmetric deployment (making life easy),
  - Horizontal scaling (no CPU bottlenecks),
  - Fault tolerance (no single-point-of-failure),
  - Rolling bug-fix updates (no downtimes).

HA dCache: SRM

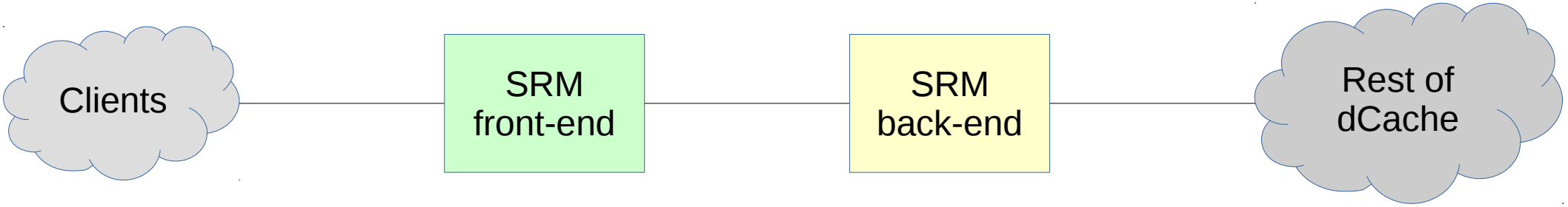
- **Split** the GSI "front-end" from "SRM engine"
- Allow **multiple front-ends**:

  horizontal scaling for encryption overhead
- Allow **multiple back-end "SRM engines"**:

  each scheduled request is processed by the same SRM engine, load-balancing and fault-survival.
- Support for **HAProxy protocol**

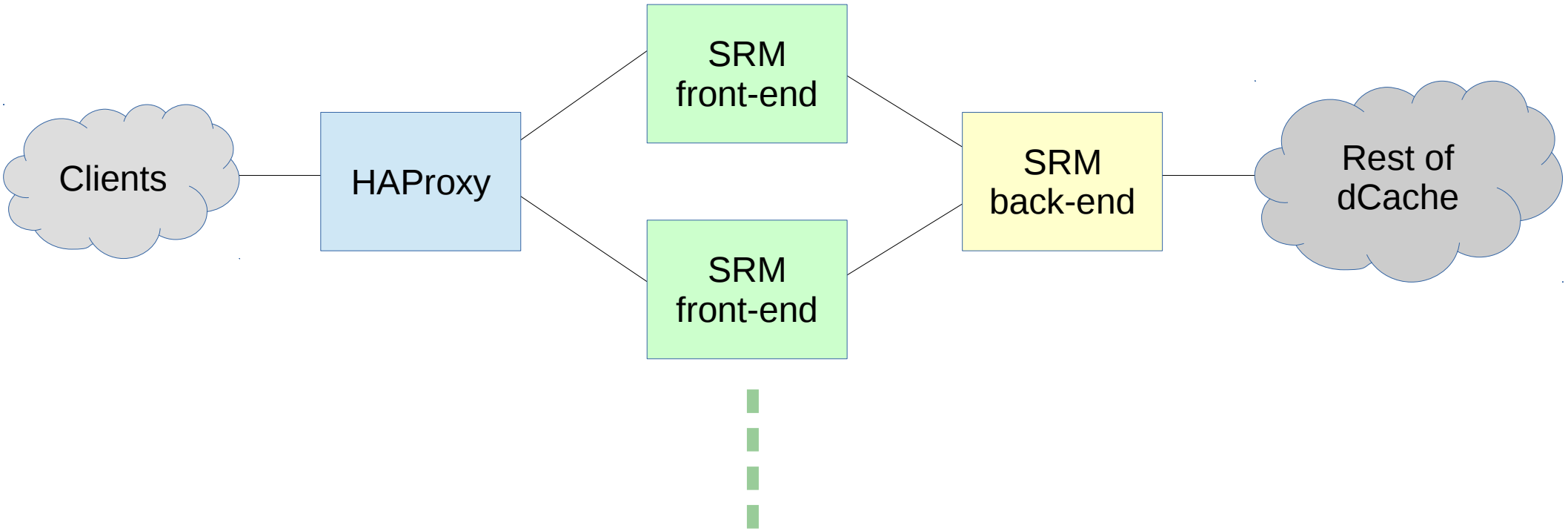
  using TCP mode, rather than HTTP mode.

Pencil sketch of possible deployment



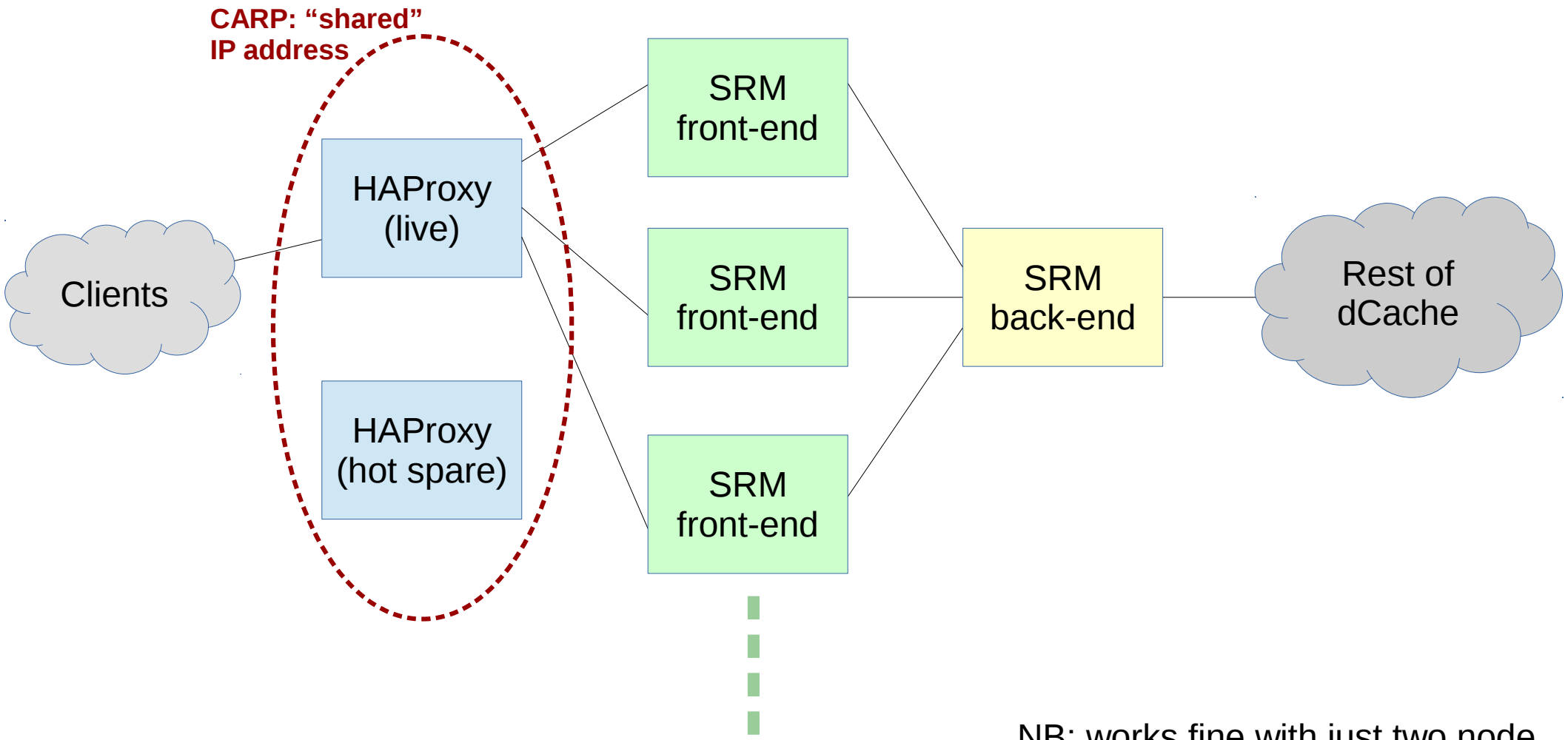
NB: works fine with just two node

Pencil sketch of possible deployment



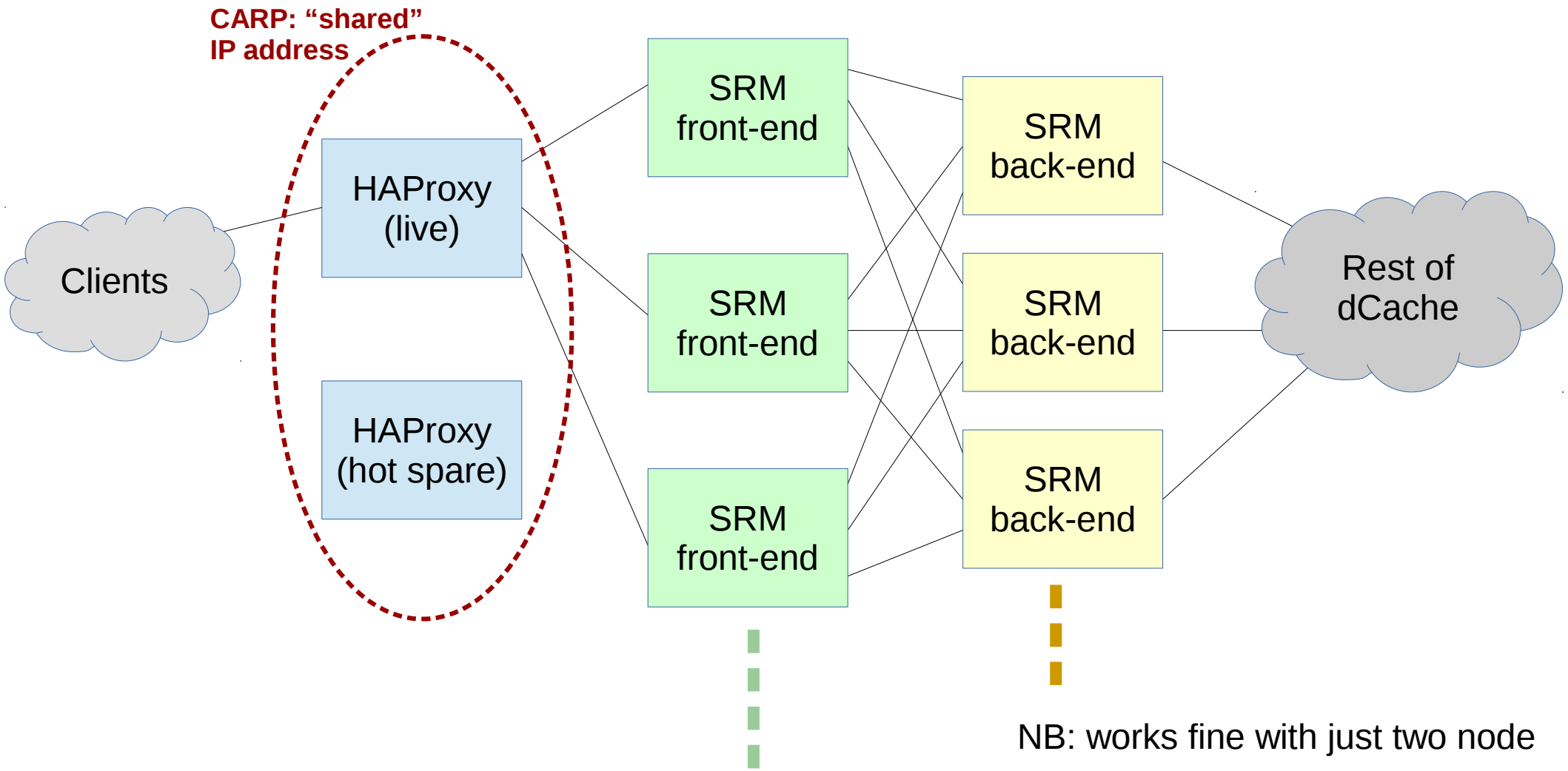
NB: works fine with just two node

Pencil sketch of possible deployment



NB: works fine with just two node

Pencil sketch of possible deployment



NB: works fine with just two node

HA dCache: general protocol remarks

- Should work fine for **TLS-based** protocols (SRM, gsiftp, webdav, gsidcap)

  Needs **load-balancer hostname** as a Subject Alternate Name (SAN) in the X.509 certificate

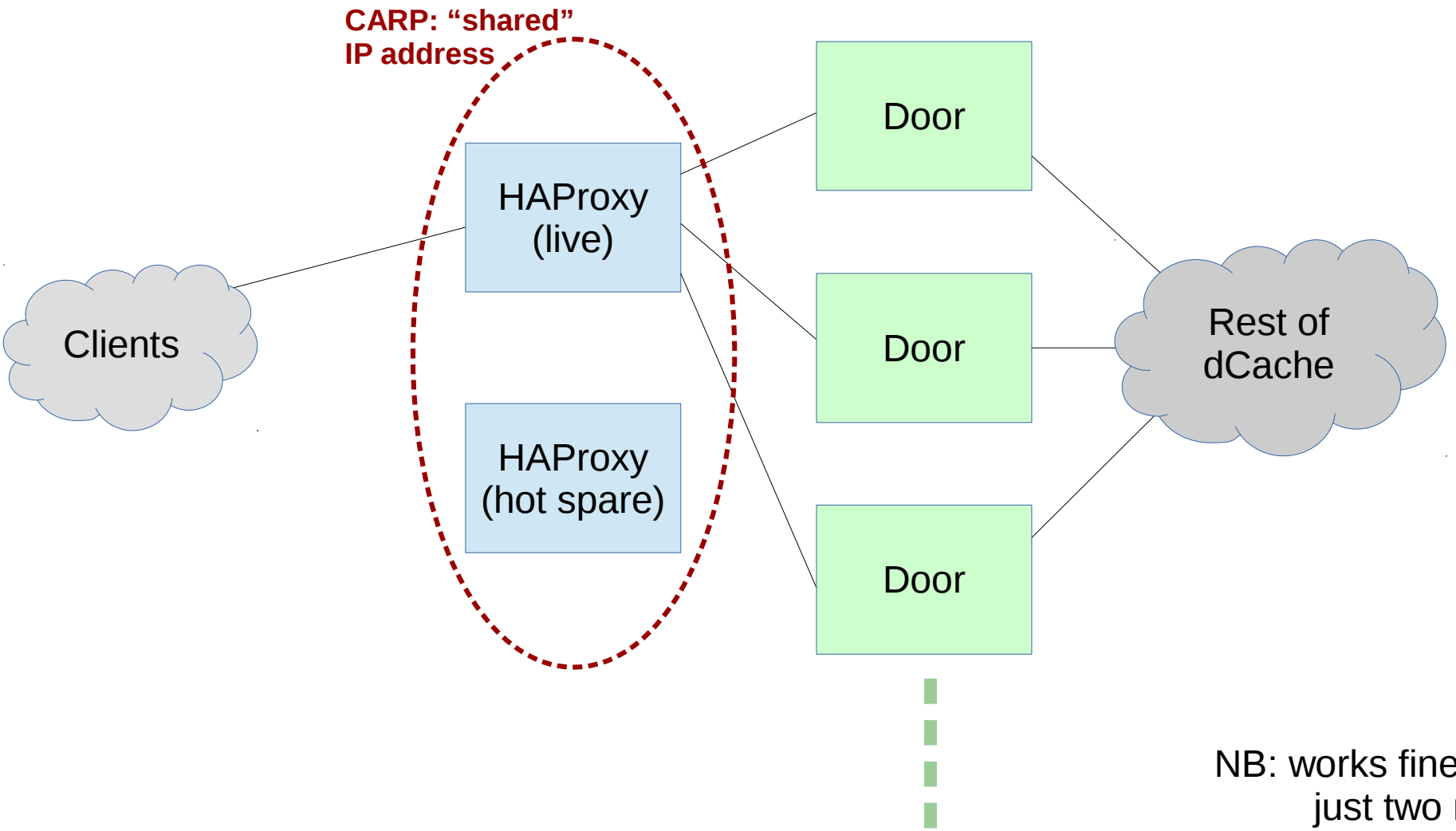
- Can have SRM redirects clients to individual doors, rather than using HA proxy:

  SRM already provides load-balancing.

- HAProxy protocol used to discover **client IP address**:

  de facto industry standard.

Pencil sketch of possible deployment



NB: works fine with just two node

HA dCache: FTP

- Updated to understand **HAProxy protocol**.
- **IPv4 and IPv6** supported.
- **Data channels** connect directly to pool or door, bypassing HAProxy.

HA dCache: other protocols

- **WebDAV**: nothing major needed
- **xrootd**: updated to understand HAProxy protocol. As usual so-called "GSI" xrootd sucks:
  - special care needed over x.509 certificate
  - kXR_locate returns IP address; makes host name verification hard.
- **dcap**: updated to understand HAProxy protocol; No other major changes needed.
- **NFS**: not updated to support HA.

HA-dCache: status and next steps

- **More details** presented to dCache admins:

  dCache workshop and "dCache Presents..." live webinar.

- Received **considerable interest** from sites.
- Deployed in **production** at NDGF

  Rolling out HA deployment to catch bugs

Other thoughts/issues on data mngmt

- Deleting until enough free capacity:

  feedback loop with delay is **unstable algorithm**!

- **Concurrent uploads** of the same file:

  Seen many times "in the wild" (ATLAS, CMS, …)

  SRM mostly protects us from this (except for "FTS srmRm bug"!)

  Not clear what will happen if not using SRM?

- **RFC 3310** HTTP checksums: supported
- **MD5 & ADLER32**: both work, but no dynamic calculation.
- **RFC 4331** WebDAV quota support:

  Work started, anticipate being in dCache v3.0.

SRM reflections

- We (dCache.org) are **NOT abandoning SRM**:
  - We have invested heavily in cleaning- and speeding it up.
  - New client release, including **srmfs** an interactive SRM shell.
- **It works** – why replace a working system?

  By now the spec and implementations are **well understood**.
- It has several unique features that would need to be re-implemented (e.g., see RFC-4331) – **wasting effort**.
- Biggest downside is NOT the protocol but the bindings & clients – this is fixable.
- Certainly, declaring SRM dead is a **self-fulfilling prophesy**.

Backup slides