HEPIX VWG Image transfer.

HEPIX virtulisation working group : 6 Month update

A short summary of the Documentation produced.
HTML PDF A4 PDF Letter

Owen Synge
HEPIX VWG Image transfer.
HEPIX fall 2011

HELMHOLTZ ASSOCIATION

DESY

HEPIX VWG Assumptions

- For new customers Cloud may be all we need.
  - But HEP is not a new customer.
- HEP Experiment software is currently partially trusted.
  - Sites allow NFS 3, sites are not ready for untrusted images.
  - NFS 4.1 with Kerberous security will be practical and possible.
    - X509 to Kerberous account mappers exist in the grid.
  - HEP experiments are not yet ready to abandon rshell style data access.
- Virtualising Worker node can be transparent for grid users.
  - We need to trust images more than with a cloud infrastructure.
- Grid/Batch Queues have high efficiency and high use.
  - We should demonstrate our ideas work with the grid.
- Accountancy is mature in the Grid.
  - Clouds are working on this.

HVWG Failures 1.

- We have not yet put enough effort into understanding the differences in cloud products.
  - Planning the mechanism of software evaluation.
  - Software.
    - WN on demand from INFN.
    - Stratus Lab marketplace.
    - Stratus Lab Open nebula.
    - OpenStack.
    - Big Grid.
    - Nimbus Infrastructure.
  - Evaluation
    - Logging.
    - Stability.
    - Scalability.
    - User management.
    - Security.
    - Accountancy.
- Even planning comparison evaluation criteria is a lot of work.

HVWG Failures 2. We need to meet more often.

- We have not organized enough meetings without Tony.
- Tony proposes a face to face soon.
- The work days where a success.
  - Not enough sites can as yet host VM's in a production like way.
    - Expect this to change.
  - Stratus lab Market place seems to work.
  - Image list based transfer is working.
  - I think we need to get back to keeping minutes.
- A lot of cloud efforts are occurring.
  - We cant all even find all of them.

HVWG Failures 3. Clearing up misunderstandings

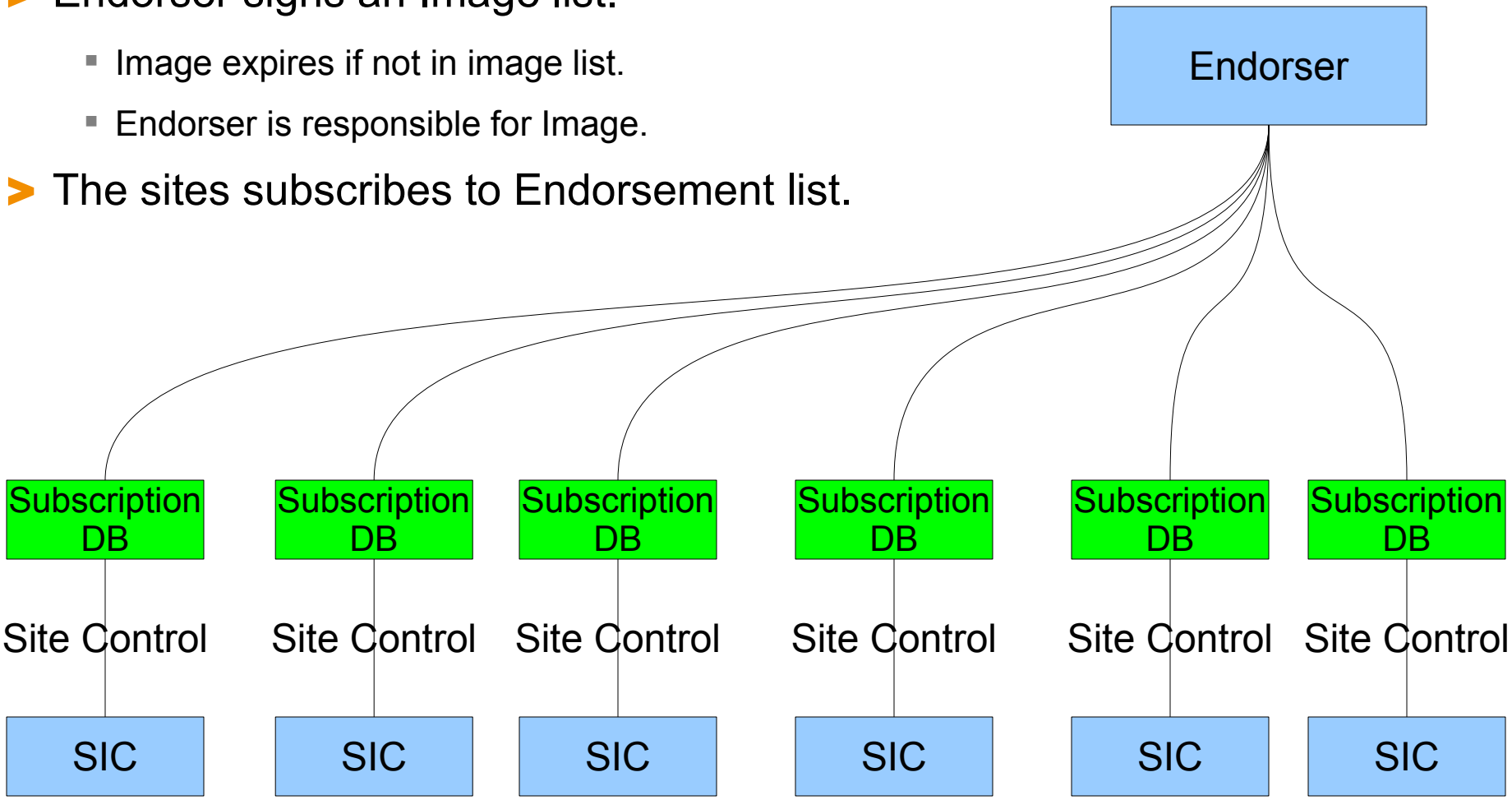
- Our work in HEPIX is intended to make it easy for scientists too.
  - Security is about enabling options (like signing an image list).
    - We welcome VO's to contribute VM for us to cache.
  - Providing guidelines how to get your image to resources.
  - Providing guidelines in making images site neutral.
- Some sites will not need to trust the VM Image.
  - User communities that access storage using secure protocols.
  - Sites with VLAN per user.
    - Giving users safe network isolation from each other.
- Some sites maybe working on alternatives to the grid accountancy.
  - Interesting work going on in big grid.
- If you have a different view how we carry on with this effort.
  - We have some experience but clouds and networks are evolving fast.
    - Help us by explaining how we should move forward.

Four Areas of Focus

- Security Policies.
  - Latest Draft
    - https://documents.egi.eu/public/ShowDocument?docid=771
  - “security-related policy requirements for the generation, distribution and operation of virtual machine images”
    - Policies are Valid even for Secured by VLAN systems.
- Image Creation.
  - How to make an image XEN KVM neutral.
- Image Transfer.
  - Most of this talk.
- Image Contextualization
  - Attach a ISO image (virtual CDROM) of site specific batch queue client software.
  - Boot time scripts call ISO image.

HVWG model is Publish and Subscribe RSS.

- Endorser signs an Image list.
  - Image expires if not in image list.
  - Endorser is responsible for Image.
- The sites subscribes to Endorsement list.



Who Publishes/Subscribes?

- CernVM.
  - 28 images for different VM formats.
  - 12 months expire time.
- Stratus Market place.
- UVIC.
  - 1 month lifetime.
- CERN
  - Instantiated images from CernVM, UVIC.
- Edinburgh
  - Instantiated images from via VMILS.
  - Instantiated images from the Stratus Marketplace, CernVM,UVIC
- INFN
  - Instantiated images via VMLIS on SL 5.

DESY

Image security.

- Image to Meta data binding.
  - Cryptographic hashes.
    - It is easy to compute the hash value for any given data.
    - It is infeasible to generate a message that has a given hash.
    - It is infeasible to modify a message without hash being changed.
    - It is infeasible to find two different messages with the same hash.
  - Chose to use sha512 and file size to validate data.
    - Following Stratuslabs recommendation.
  - Other hashes can be added.
    - If sha512 and size are later found to be too week.
  - URI to retrieve image.
    - Can be cached locally.
  - Each image has a UUID
    - So we know which image is expired and which is upgraded.

Meta-data Security.

- Meta-data authenticity.
  - X509 + signatures. (SMIME or XML signatures)
    - Gives non repudiation, and confidence in who endorsed.
    - Give tamper proof message.
    - Signature can be checked by all clients,
    - Allows checking of historic meta-data changes.
  - Version number.
    - Prevents man in middle attacks.
    - Man In Middle attempts to return an old list blocked by this.
  - UUID on Image and Image list
    - Allows messages to be identified.
    - So messages cannot effect each other.
    - So images can be expired and updated.

Example of what signed meta-data can look like


```
------EAE3006C97F670EE450F46AC8DF4C070
{
   "dc:date:created": "2011-03-10T17:09:12Z",
   "dc:date:expires": "2011-04-07T17:09:12Z",
   "dc:description": "a README example of an image list",
   "dc:identifier": "4e186b44-2c64-40ea-97d5-e9e5c0bce059",
   "dc:source": "example.org",
   "dc:title": "README example",
   "hv:endorser": {
      "hv:x509": {
         "dc:creator": "Owen Synge",
         "hv:ca": "/C=DE/O=GermanGrid/CN=GridKa-CA",
         "hv:dn": "/C=DE/O=GermanGrid/OU=DESY/CN=Owen Synge",
         "hv:email": "owen.synge@desy.de"
      }
   },
   "hv:images": [
      {
         "hv:image": {
            "dc:description": "This is an README example VM",
            "dc:identifier": "488dcdc4-9ab1-4fc8-a7ba-b7a5428ecb3d",
            "dc:title": "README example VM",
            "hv:hypervisor": "kvm",
            "hv:size": 2147483648,
            "hv:uri": "http://example.org/example-image.img",
            "hv:version": "1",
            "sl:arch": "x86_64",
            "sl:checksum:sha512":
"8b4c269a60da1061b434b696c4a89293bea847b66bd8ba486a914d4209df651193ee8d454f8231840b7500fab6740620c7111d9a
17d08b743133dc393ba2c0d4",
            "sl:comments": "Vanila install with contextulization scripts",
            "sl:os": "Linux",
            "sl:osversion": "SL 5.5"
         }
      }
   ],
   "hv:uri": "http://example.org/example-image-list.image_list",
   "hv:version": "1"
}
------EAE3006C97F670EE450F46AC8DF4C070
Content-Type: application/pkcs7-signature; name="smime.p7s"
Content-Transfer-Encoding: base64
Content-Disposition: attachment; filename="smime.p7s"

MIIHdAYJKoZIhvcNAQcCoIIHZTCCB2ECAQExCzAJBgUrDgMCGgUAMAsGCSqGSIb3
DQEHAaCCBSUwggUhMIIECaADAgECAgIz7DANBgkqhkiG9w0BAQUFADA2MQswCQYD
VQQGEwJERTETMBEGA1UEChMKR2VybWFuR3JpZDESMBAGA1UEAxMJR3JpZEthLUNB
MB4XDTExMDExMDE1MDMxN1oXDTEyMDIwOTE1MDMxN1owRjELMAkGA1UEBhMCREUx
EzARBgNVBAoTCkdlcm1hbkdyaWQxDTALBgNVBAsTBERFU1kxEzARBgNVBAMTCk93
ZW4gU3luZ2UwggEiMA0GCSqGSIb3DQEBAQUAA4IBDwAwggEKAoIBAQCkgbPFZrVL
pmwf7GKBBFkwTK5V7RmlupsU3Z3FqdfMnJGn2NrrnHIhthUTCTq4WbLIZTbOEh0n
JqZgZBvYcwJV4V9pais4YlsEug+JLMbB9hZ6e2XgdjXWgLqz6vBSIf6KXi4KhCxe
a4FylvIk7OtY+bg0mg5IFHib6uP7fXhFKdBEapoi+B05wpluBMA+2DBdSt+rjzA8
SwiHUuan60VIyJAxammyOe3IKSpwyBXkQ10XjIhIpoSavqYXJboFOVzUcqxawdbX
Con2W8QfiwFKYupohG/VTusDXFT2MP4k+KxG3/rTTPWUDJme7VUPv3+CTcEO+z4v
X8/XhI44oAXIAgMBAAGjggInMIICIzAMBgNVHRMBAf8EAjAAMA4GA1UdDwEB/wQE
AwIE8DAdBgNVHQ4EFgQUGakUy66kgvuINBIf18WBXjGolqYwXgYDVR0jBFcwVYAU
xnXJKKzRC/w8/7m1HtNfO4BiEjShOqQ4MDYxCzAJBgNVBAYTAkRFMRMwEQYDVQQK
EwpHZXJtYW5HcmlkMRIwEAYDVQQDEwlHcmlkS2EtQ0GCAQAwHQYDVR0RBBYwFIES
b3dlbi5TeW5nZUBkZXN5LmRIMB8GA1UdEgQYMBaBFGdyaWRrYS1jYUBpd3luZnpr
LmRlMDUGA1UdHwQuMCwwKqAooCaGJGh0dHA6Ly9ncmlkLmZ6ay5kZS9jYS9ncmlk
a2EtY3JsLmRlcjAaBgNVHSAEEzARMA8GDSsGAQQBIDarLAEBAQUwEQYJYIZIAYb4
QgEBBAQDAgWgME4GCWCGSAGG+EIBDQRBFj9DZXJ0aWZpY2F0ZSBpc3N1ZWQgdW5k
ZXlgQ1AvQ1BTIHYuIDEuNSBhdCBodHRwOi8vZ3JpZC5memsuZGUvY2EwJAYJYIZI
AYb4QgECBBcWFWh0dHA6Ly9ncmlkLmZ6ay5kZS9jYTAzBglghkgBhvhCAQgEJhYk
aHR0cDovL2dyaWQuZnprLmRlL2NhL2dyaWRrYS1jcHMucGRmMDMGCWCGSAGG+EIB
AwQmFiRodHRwOi8vZ3JpZC5memsuZGUvY2EvZ3JpZGthLWNybC5kZXIwDQYJKoZI
hvcNAQEFBQADggEBAMbn91TOQ6r4D/aKwgIFXiXe40B7iccz/P5pCFSi1R6IC3KH
Ui4s/f9iAGl9rA21h8QAaRaJ/h1OQNlgMZbc9jDCWcqxr8wQTYAQDiBkspLT68ZO
5xVFRiq3HjkkhwnFfFzsNSiLFYZTRjChPluclYG3TEvSg8dz9Lv/IEJxE5C5IZ2d
e3CSu0vcD0DESiu/sVqPOOHi8NL/59U2ine3z23Y+piCabQCxjT0inT2MmR8UNDF
ij2JJYxlt56U/SQCEe0304w3x1jIg8vcpm4dfh+L2IjJ9hVfEeLaCyhv9Wjbmu5O
vk0yLjcEZ7b4RKeo7djVYh+5kCWJYCr/W6uGW44xgglXMIICEwIBATA8MDYxCzAJ
BgNVBAYTAkRFMRMwEQYDVQQKEwpHZXJtYW5HcmlkMRIwEAYDVQQDEwlHcmlkS2Et
Q0ECAjPsMAkGBSsOAwIaBQCggbEwGAYJKoZIhvcNAQkDMQsGCSqGSIb3DQEHATAc
BgkqhkiG9w0BCQUxDxcNMTEwMzEwMTczMzU1WjAjBgkqhkiG9w0BCQQxFgQUd43y
VT05Zk+7acFF+EeqExNI57cwUgYJKoZIhvcNAQkPMUUwQzAKBggqhkiG9w0DBzAO
BggqhkiG9w0DAgICAIAwDQYIKoZIhvcNAwICAUAwBwYFKw4DAgcwDQYIKoZIhvcN
AwICASgwDQYJKoZIhvcNAQEBBQAEggEAkA0RgB5AkGIYvFsFETzx7QHKWu9qas5k
vIHn2a+EpRE9K1p+qrFNzS53E2BGqubyRcePfgG/WyGqYOK2h20d6GZH+ENUFkvM
EAthbvQaHye6WEvF/0GUrr0QUBT1gQswkkryPHcqTVmJANQORakkNvCwynEBmfSC
vb2TEppRuOCmxx3zqrzMr7zPNPY4w2+YaXQ1fHfmEmOrlf0ImP20TyTKIoQWqzbq
WXwlRhZBUoD9zfiEM/jFvOvkuxLkQeiEcSzlLAGHXsHJ3anPMX9sobJFbJl0wYdN
sUOInHRhksokh2ow68KZK4vXLI173v5yZE7FZZ1Gl9T+YpkmOlW4iQ==

------EAE3006C97F670EE450F46AC8DF4C070--
```


Please Note:

XML implementation would look different but will be functional identical.

XML signatures store signature as XML.

Meta-data subscription DB

- Mostly no admin interaction!
  - Database is auto created.
  - All subscriptions updated from a cron script.
  - All data is derived from subscriptions to image lists.
    - So just need to store signed image lists which you should anyway.
  - Migration is simply install a second in parallel.
- Simple RDBMS
  - Almost No critical data to back up.
    - You might want to back up users and subscription list.
- Image cache as a client of the subscription data base.
  - 3 Directories (should be on same file system).
    - Current, Expired, and Downloading directories.
  - Current images are be validated.
  - JSON catalog for each directory.

DESY

Whats been done in the last 6 months.

- HVWG Document generated nightly from DESY SVN.
  - HTML PDF A4 PDF Letter
  - http://grid.desy.de/vm/hepix/vwg/doc/
- Completed the Image cache code.
  - Initial version subscribed to all images in an image list.
- New feature: Images can be selected to be cached.
  - CERNVM has 28 different versions of same image.
    - Some of these where about 30G and not wanted.
- Test suite had to be developed.
  - Bugs fixes where getting to much work to replicate.
- Image transfer software is now stable and tested.
  - 2 data base schema evolutions since last meeting.

Whats been done in the last 6 months.

- Bug Fix: Subscription software now enforces UUID constraints.
  - Warn admin when unregistered images are added to a subscribed list.
  - Ignore image UUID's that clash with existing image UUID's.
- Enhancement: Logging has been upgraded
  - Now use standard python logging configuration.
    - Allowing admins to change logging.
- New Feature: Publisher User management enhanced in subscriber code.
  - Can now add delete user separately from subscriptions.
  - Can bind users to one or more subscription.
  - NIKEF feedback wanted changes to subscription model.
    - Not just based on URL, but on approved DN.
    - Now implemented.
- New Feature: Support Host certificates.
  - Should I add this feature?

DESY

What we learn through subscription and publishing.

- Code to publish image lists is almost trivial.
  - Minimal less than 70 lines of python (including the metadata and white space)
  - Command line code is considerably bigger as its easier to use.
- Subscribing to image lists is more complex than publishing.
  - Test suites have been needed.
- Test Image list updates are infrequent.
  - Our test image list providers are not updating regularly.
  - Test suites give confidence that updates are handled.
  - Hope this will change with more production VM hosting environments

Whats still to look at (so far)

- Network technology. (VLAN, subnets, monitoring at only network level)
- Contextualization
  - Can we just use the open nebular model?
    - Was our first plan.
  - Device drivers change with VHost setup.
    - First drive can be: sda,hda,vda
  - CDROM ISO mounting.
  - Need to come up with more conventions.
- Clouds and fair share.
  - UVIC cloud scheduler etc.
- Pilot factories and Virtualisation.
  - Can we just launch precooked images that hook to the pilot frameworks.
- Finalise our first iteration of documents.
- Stratus Lab's Image marketplace.
  - Need to look into this further. (many bug fixes since first looks)

DESY

Last words : Coming to the End of first Iteration.

- VMIC is replacing VMIC Strawman.
- HEPix Virtualisation Working Group is coming to end of first iteration.
  - We did what we where created for but its clear more work is needed.
  - We documented what we did.
    - http://grid.desy.de/vm/hepix/vwg/doc/html/index.shtml
    - http://grid.desy.de/vm/hepix/vwg/doc/pdf/Book-a4.pdf
    - http://grid.desy.de/vm/hepix/vwg/doc/pdf/Book-letter.pdf
- Glue supports Virtualised Execution Environments.
  - Should we start using this?
- I hope for changes to Cream CE to allow users to select images.
  - JDL Change/Addition?
  - We need support in Batch Queue integration.
- Image creation and Contextualization has only been demonstrated.
- Image List base VM Image caching is demonstrated.
  - Progress to putting it in EPEL.

DESY

Summary (Concepts matter not implementation)

- Policy for dealing with Virtual Machine Images.
- Recipes for image creation as part of a Virtualised Grid Worker Node.
- Signed image lists define images that are endorsed.
  - First version of meta data is defined.
  - Non repudiation of image lists through signatures.
- Only Images on current Image list are endorsed.
  - This means images expire when not in current image list.
- Principle is generic to Clouds, Virtualised Worker Node.
- Implementation of Message Generation/Subscription exist.
  - Working on getting release into EPEL repository.
  - Further working starting in CERN/Academia Sinica
- We recommend the concept of Signed image lists.
  - And using Publish Subscribe model.

DESY

DESY 1 Cloud : Network View : Current + Planning

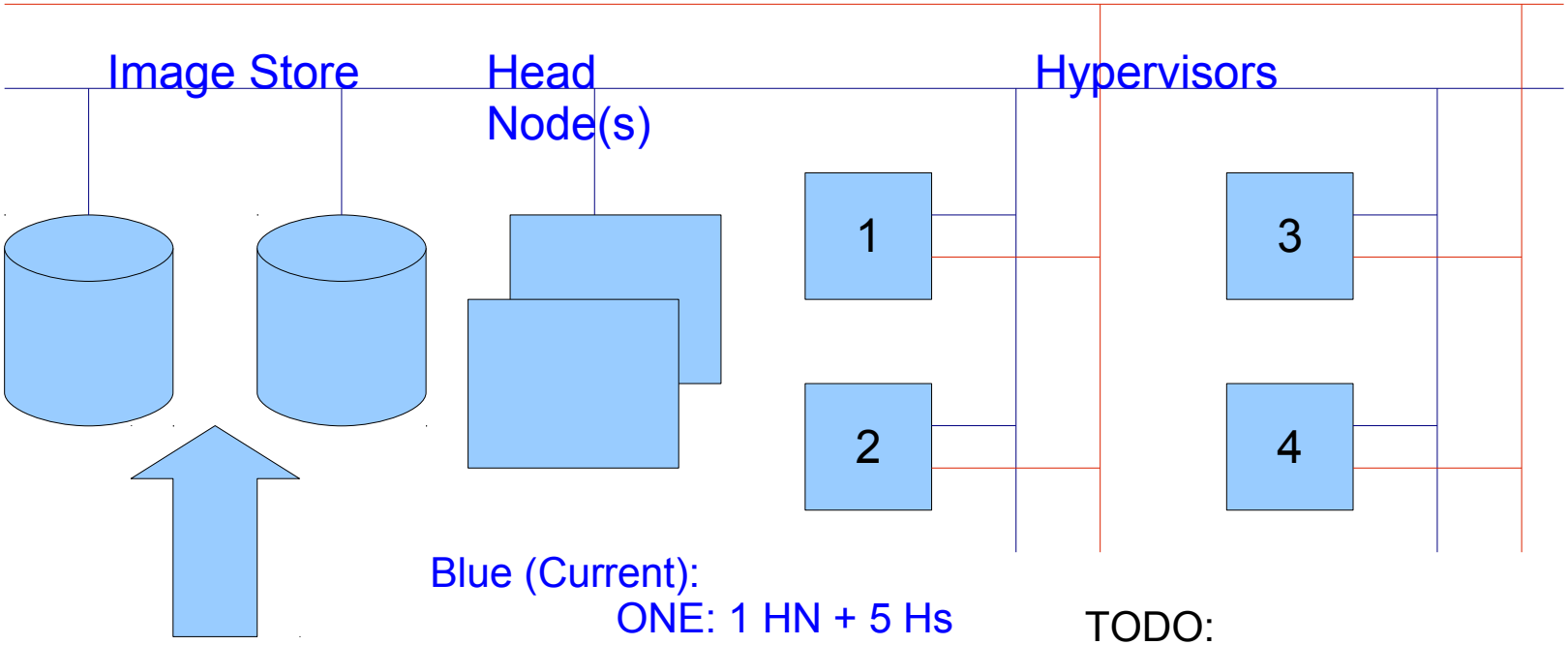


Image importing

VMILS

Blue (Current):
ONE: 1 HN + 5 Hs
Cloud Layer
256 private IPs
Planned:
As Cloud Admin Net

with
ONE: 2 Hns + 14 Hs

Red(Planned):
VLAN's (802.1Q +):
512 private Cloud IPs
512 public Cloud IPs
nnn Selected DESY

TODO:
High availability Head nodes
Storage is not yet fixed

DESY2 : DISH

- Desy ImageSHaring : Reaching pre production soon.
- Running on dCache WebDav Door.
- Using our Quattor to publish a worker node.
- Plan to publish fresh images weekly.