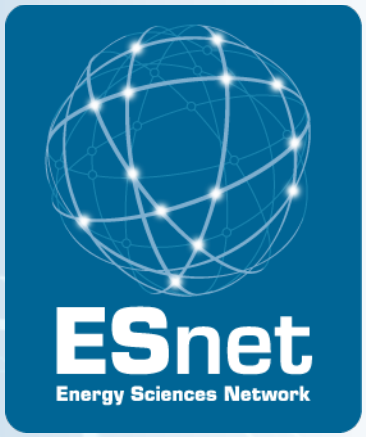


Feeling the Brady Bunch's Pain

Michael Sinatra, Network Engineer

ESnet Network Engineering Group

NANOG 58

New Orleans, LA

June 2013

Overview

Beginning state: 2 networks, 3 platforms, 3 routing protocols

Guiding principles

Merging the IGP in ESnet4

Merging the routing from ESnet4 & ANI Prototype to ESnet5

Lessons learned and conclusions

Here's the story…

From mid-2011 to late 2012, ESnet operated two networks: ESnet4 and the ANI Prototype Network.

- ESnet4: ESnet's main production network; this incarnation of ESnet had been in operation since 2007.
- Had separate links for regular IP traffic and for large-scale science data flows managed by OSCARS, Esnet's dynamic circuit provisioning software. This latter sub-network was known as "SDN"--for "Science Data Network," not "Software Defined Network."
- ANI: Advanced Networking Initiative Prototype Network: The prototype 100G network using Internet2/ESnet optical transport infrastructure. "… this multi-year, $62 million Recovery Act investment is creating a blazingly fast prototype network of unprecedented capacity that will stimulate not just science, but the development of tomorrow's networking technologies." (http://www.es.net/news-and-publications/press-kit/ani-high-speed-network-for-national-scientific-competitiveness/)

Here's the story…

- The two networks were logically separate.
  - Both operated by ESnet.
  - No common IGP.
  - Each network had separate ASN and IP addressing.
  - Networks peered with each other using eBGP.
  - Each network also used separate iBGP meshes to carry non-backbone routes through each backbone.
  - No iBGP between ESnet4 and ANI Prototype (only eBGP).

→ The goal was to take these two networks and merge them into a single, 100G-capable production network, with no network-wide partitions or interruptions.

Here's the story…

	ESnet4	ANI Prototype
Routing platforms	Cisco (7200), Juniper (M, MX, T)	Alcatel-Lucent (ALU) 7750SR
IGP	IS-IS ST (IPv6) OSPFv2 (IPv4)	IS-IS MT (IPv4 and IPv6)
ASN	293	3427
Science Data traffic vs. traditional IP traffic	Separate "IP" and "SDN" links and routers	Converged
Backbone addresses	main ESnet backbone /16	separate swamp /24
Optical transport	Level(3) wave service	ESnet-I2 Managed Ciena 6500 over L(3) fiber

Guiding principles

- Obviously, the backbone cannot go down completely.
- The backbone also cannot be partitioned (i.e. traffic traversing a portion of the backbone gets blackholed due to no route/bad routes).
- Treat IPv6 the same (as important) as IPv4. "SLA parity" between protocols.
- Minimize or preferably prevent/eliminate any site outages.
- Minimize platforms in ESnet5 (i.e. get rid of one of the three platforms)
  - Before or after transition?
  - Ideally before, but there simply wasn't time. Removing "third platform" still a priority.

Merging the IGP

- As you can see from the starting point table, ESnet4 used IS-IS as the IPv6 IGP and OSPFv2 as the IPv4 IGP.
  - iBGP and eBGP were both multi-protocol and handled both v4 and v6, although eBGP uses separate v4 and v6 peerings.
  - OSPFv2 had originally replaced EIGRP (yeah, it has been around that long); IS-IS had been used since at least the early 2000s for IPv6 routing.
  - The plan was always to merge IPv4 routing into IS-IS.
- Decision had to be made as to whether to merge IGP into IS-IS before, during, or after the ESnet5 transition, or to not merge at all.
- → Decision was made to merge before the transition.

Merging the IGP

- Another option was to keep OSPFv2 and merge IS-IS into OSPFv3.
- Pro-merge:
  - IS-IS is frequently run among ISPs, as well as R&E NRENs and RONs. It is well-known within the service provider community.
  - Single protocol easier to troubleshoot (at least in certain circumstances), maintain.
  - ANI network already using IS-IS only.
- Con-merge:
  - More work to do, more surgery on the patient as we're also merging networks.
  - There were some advantages to OSPFv2-v3 as we would see later. But IS-IS→OSPFv3 would have been additional work as well.

Merging the IGP

- Okay, we're doing IS-IS. Do we do single-topology or multi-topology?
  - Pro-Single-topology:
    - Easier to maintain. Only one set of metrics, less complicated protocol.
    - Easier to migrate. Basically, it's already done; we just need to change filters to start importing IPv4 routes.
  - Pro-Multi-topology:
    - Single-topology can black-hole traffic if one address family is removed from an interface. (Thinking about the day when we may have IPv6-only links.)
    - Multi-topology allows for more flexibility in routing (different topologies based on address-family, different metrics, etc.
    - ANI network already using MT.
- → Multi-topology

Merging the IGP

- How do we do the switch from ST to MT?
  - Fast switch. Just turn it on.
    - Con: Easy, but we would lose IPv6 routing information while the transition was taking place. *This violates our guiding principles.*
  - Graceful switch, using IS-IS MT compatibility mode.
    - Con: Whoops, this is only supported on Cisco, not Juniper.
  - Kludgy graceful switch, using OSPFv3 as a temporary routing protocol.
    - Con: Who are we, Caltrans?
    - Pro: But it turns out that it's very easy to programmatically build OSPFv3 configurations from the existing IS-IS ones.
- → We'll do the kludgy switch. At some point we'll have OSPFv2, OSPFv3, and IS-IS MT routing IPv4 and IPv6 twice.

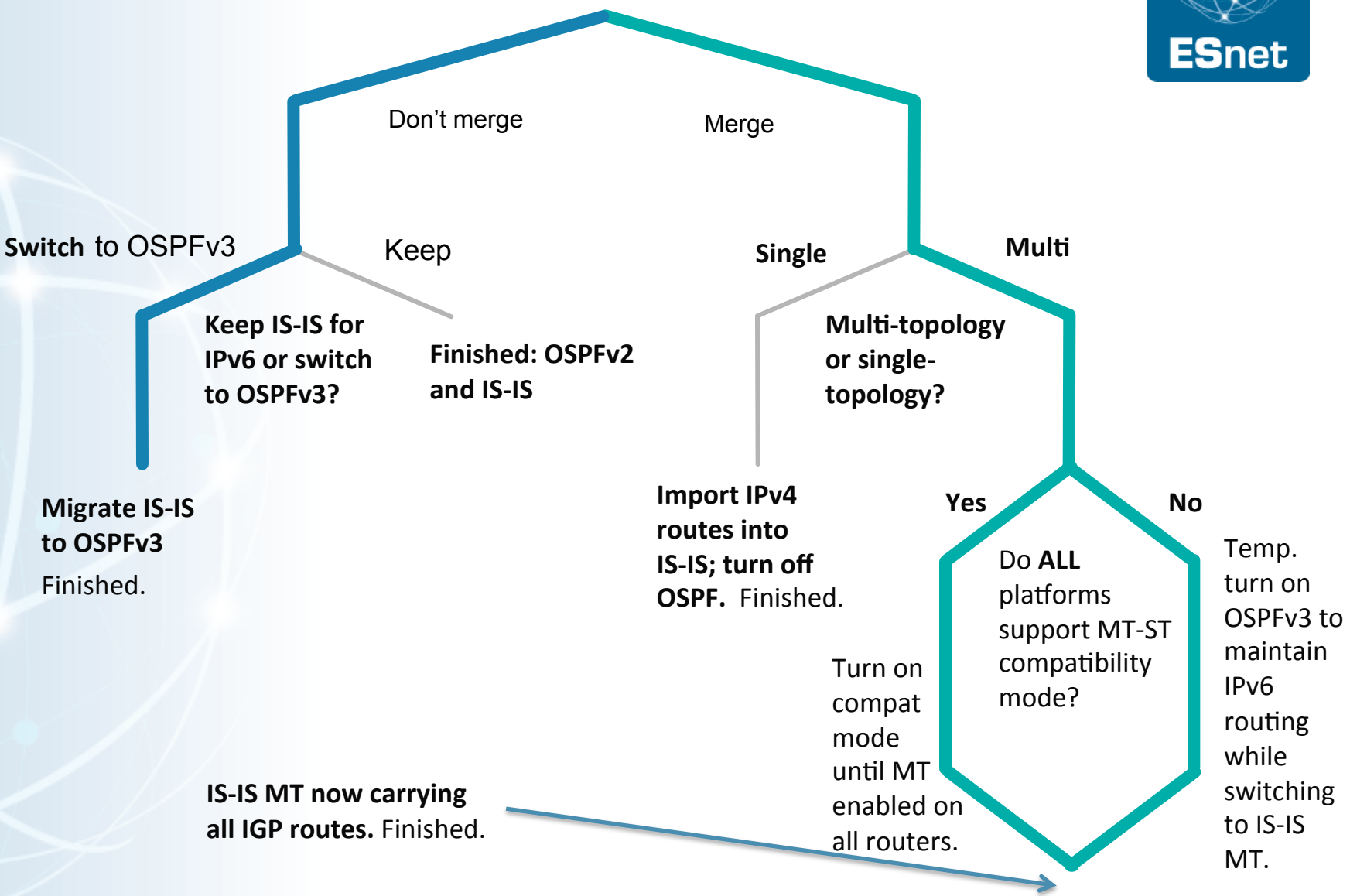
Merging the IGP

- There's one more problem and one more decision, having to do with default routes.
  - ESnet provides routing options to its sites: Full Internet tables, defaults only, or defaults+R&E routes.
  - For various reasons, announcing default in core routers is very tricky.
  - Instead we use a special default prefix, which our sites use to generate default routes at the site border.
  - This default prefixes *should* only be carried in iBGP. But at the time, they were also carried in OSPFv2.

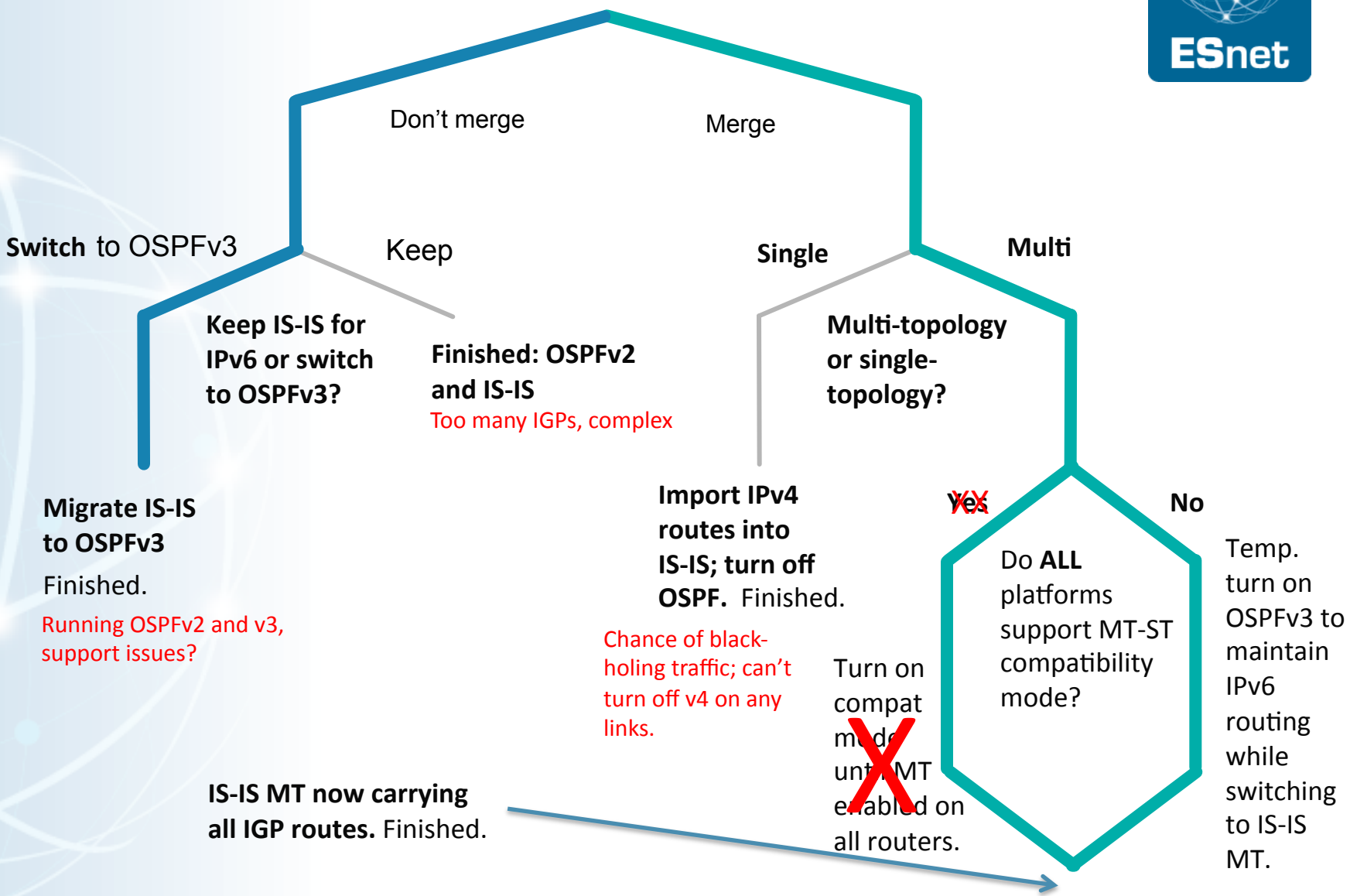
Merging the IGP

- Do we clean up the default prefix situation as part of the IGP merge or do we replicate it into IS-IS.
  - Con - Replicate:
    - It's more up-front work.
    - It's yucky.
    - It preserves a mess that needs to be cleaned up.
    - Cleaning up sooner may reduce the overall pain.
  - Con - Don't replicate and clean up:
    - Much more likely to cause problems for sites that somehow rely on getting these prefixes through the IGP
- → Let's bite the bullet and clean it up.

DECISION TREE

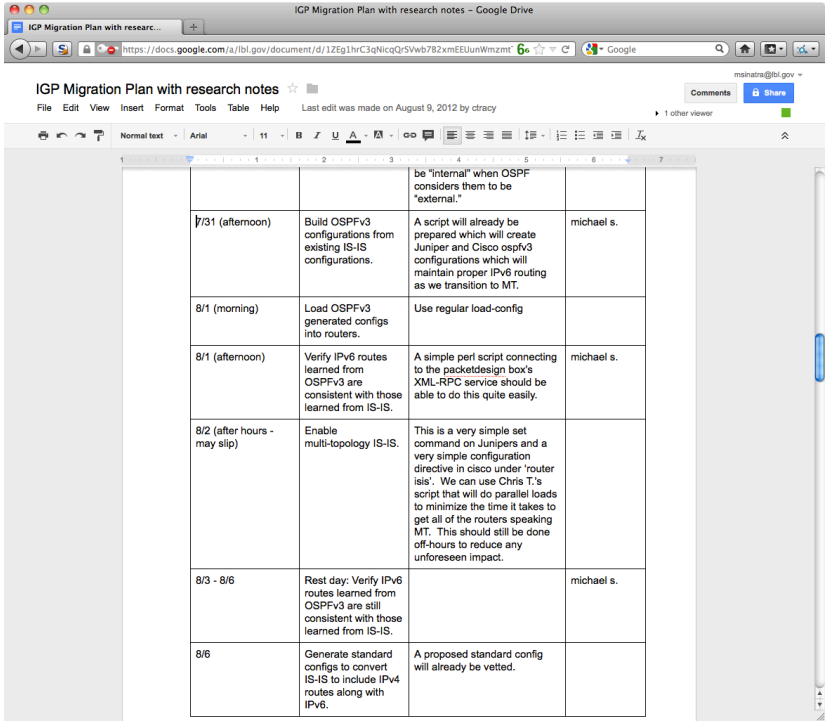


DECISION TREE
IGP MERGE



Merging the IGP: Doing it

- First, create a detailed plan with step-by-step instructions, timelines, and task assignees.
- Create any necessary configuration cut-and-paste templates (or any scripts to generate configurations).
- Do it, do it carefully, and document what you did.



		be "internal" when OSPF considers them to be "external."	
7/31 (afternoon)	Build OSPFv3 configurations from existing IS-IS configurations.	A script will already be prepared which will create Juniper and Cisco ospfv3 configurations which will maintain proper IPv6 routing as we transition to MT.	michael s.
8/1 (morning)	Load OSPFv3 generated configs into routers.	Use regular load-config	
8/1 (afternoon)	Verify IPv6 routes learned from OSPFv3 are consistent with those learned from IS-IS.	A simple perl script connecting to the packetdesign box's XML-RPC service should be able to do this quite easily.	michael s.
8/2 (after hours - may slip)	Enable multi-topology IS-IS.	This is a very simple set command on Junipers and a very simple configuration directive in cisco under 'router isis'. We can use Chris T.'s script that will do parallel loads to minimize the time it takes to get all of the routers speaking MT. This should still be done off-hours to reduce any unforeseen impact.	
8/3 - 8/6	Rest day: Verify IPv6 routes learned from OSPFv3 are still consistent with those learned from IS-IS.		michael s.
8/6	Generate standard configs to convert IS-IS to include IPv4 routes along with IPv6.	A proposed standard config will already be vetted.	

Merging the IGP: Basic steps

- Clean up any messes in IS-IS with respect to IPv6 routes.
  - Non-backbone routes being carried in IGP.
  - Default prefix issues.
- Vet end-state IS-IS config.
- Generate and load OSPFv3 configs based on IS-IS IPv6 config.
- Verify that IS-IS and OSPFv3 routes are consistent.
- Switch to IS-IS MT. Allow IPv4 backbone routes into IS-IS.
- Verify again that IS-IS and OSPFv3 routes are consistent.
- Verify that IS-IS and OSPFv2 routes are consistent.
- Turn off OSPFv3. Turn off OSPFv2.

Merging the IGP: Outcomes and Lessons

- It worked amazingly well. But in the process of the transition, we found some routes weren't getting into OSPFv3 that were in IS-IS.
  - These were due to broken IPv6 router interface configurations.
  - IS-IS didn't pick these up, but OSPFv3 did. This underscores the need for address-family BFD in IS-IS.
  - ALU has BFD for both IPv4 and IPv6, but Juniper doesn't.
  - IPv4/IPv6 feature parity rears its head again.
  - BFD not implemented yet, currently working on a test and implementation plan.

Merging the IGP: Outcomes and Lessons

- If we hadn't tried cleaning up the default prefix stuff, there would have been absolutely zero outages. However…
  - We did clean up the default prefixes so that they were only carried in iBGP not in the IGPs. This caused two sites to go offline for a while and we had to scramble to fix some non-standard Cisco configs.
  - → Wouldn't have happened if we had eliminated the Ciscos.
  - → But it's nothing intrinsic to the Ciscos—had we audited the configs and standardized them, this also wouldn't have happened. We did this on one router that was woefully out-of-date and it came through just fine.
  - We didn't have enough time to eliminate the Ciscos. We may not have even had the time to audit their configs.

Merging Routing/Routers Guiding principles (redux)

- Obviously, the backbone cannot go down completely.
- The backbone also cannot be partitioned (i.e. traffic traversing a portion of the backbone gets blackholed due to no route/bad routes).
- Treat IPv6 the same (as important) as IPv4. "SLA parity" between protocols.
- Minimize or preferably prevent/eliminate any site outages.
- Minimize platforms in ESnet5 (i.e. get rid of Ciscos)
  - Before or after transition?
  - Ideally before, but there simply wasn't time. Removing Ciscos still a priority.
- We can, however, shut down the ANI network for a period of time (days/ weeks).
- We can schedule extremely short blips in OSCARS circuits, but they must be coordinated with sites.

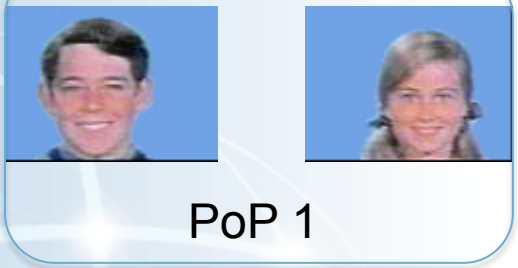
Merging the routing: Basic steps

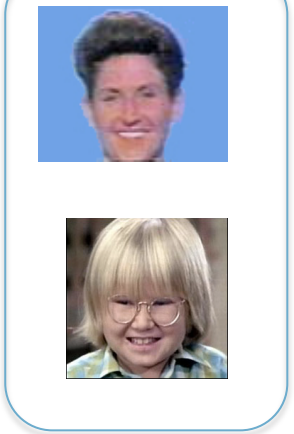
- We're adding a new platform to ESnet. We need to first set configuration standards, best practices, etc., and build templates.
- Build tools to manage the transition and maintain the network with the ALUs integrated into ESnet5. *platform-tool parity!*
- Shut down ANI network and its eBGP peerings with ESnet4.
- Assign addresses in ESnet4 backbone for new routers.
- Generate standard configs with new addresses, to participate in ESnet routes.
  - IS-IS, iBGP mesh (no external peers on the ALUs yet)
  - 100GE links initially costed out of production
  - ESnet4—ANI links were LAGs. Convert them to ECMP routed links and augment them where traffic demands

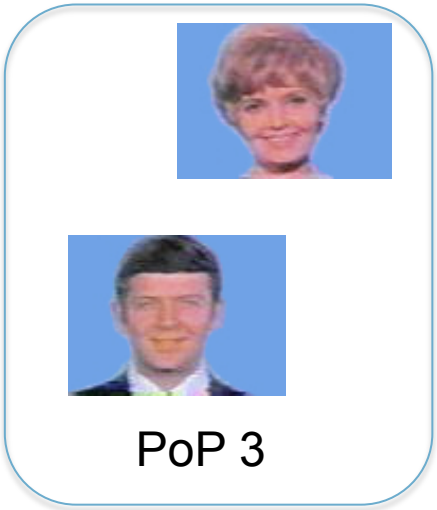
Merging the routing: Basic steps (cont)

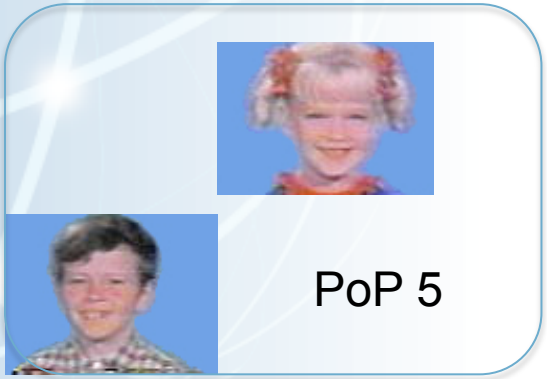
- Lower 100G links' costs and put into production.
- Raise legacy ESnet4 10G links to effectively move them out of production.
- Schedule, coordinate and move OSCARS circuits off of legacy links and onto new converged 100G links.
- Remove from production (deactivate and disable) 10G ESnet4 links. (Necessary to do this by December 2012 for contractual obligations.)
- Delete configurations and reclaim point-to-point networks.
- Delete associated DNS entries.
- ESnet5 is now in production.

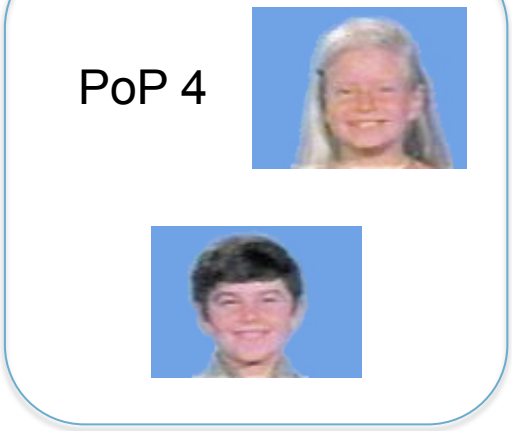
How we formed a network



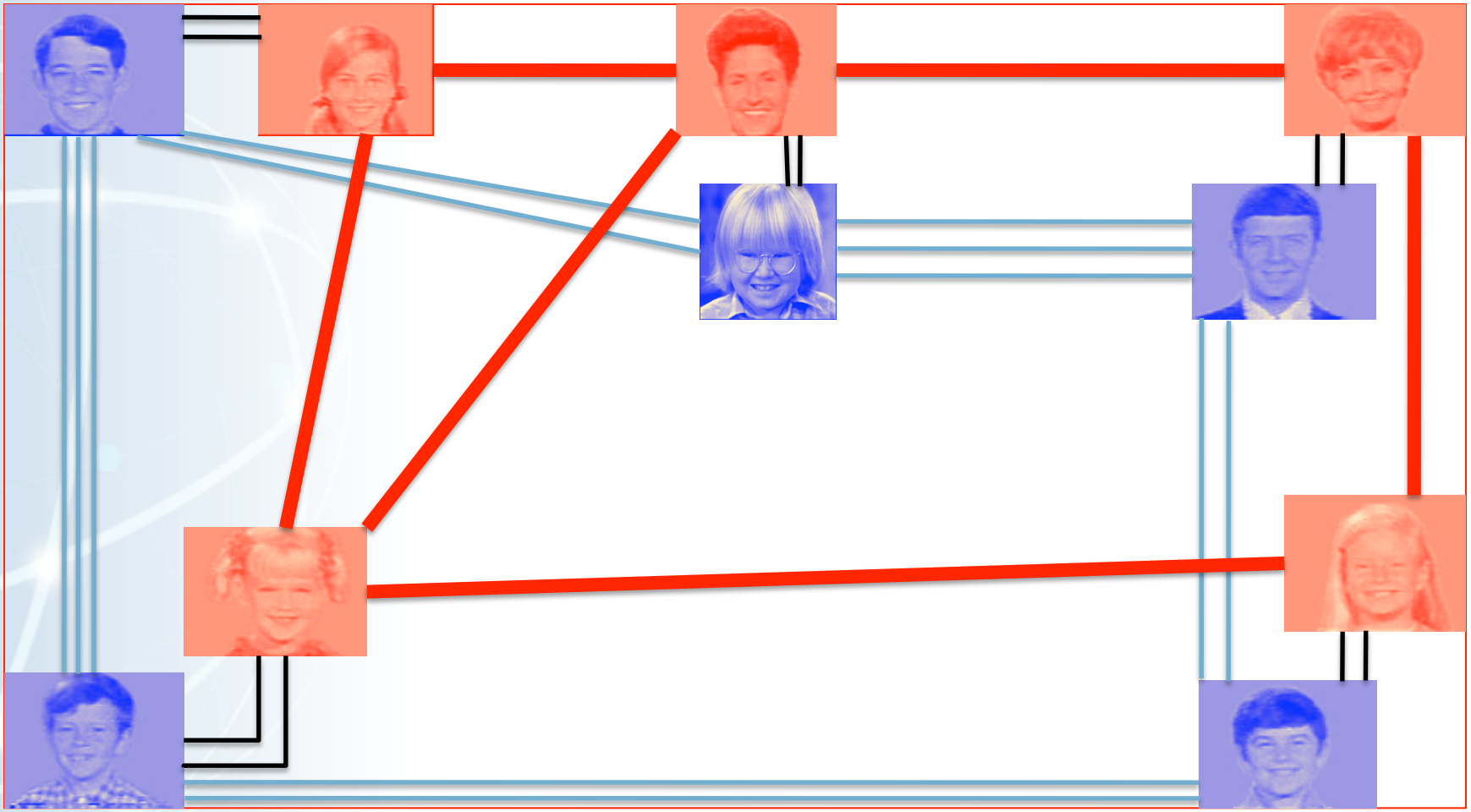




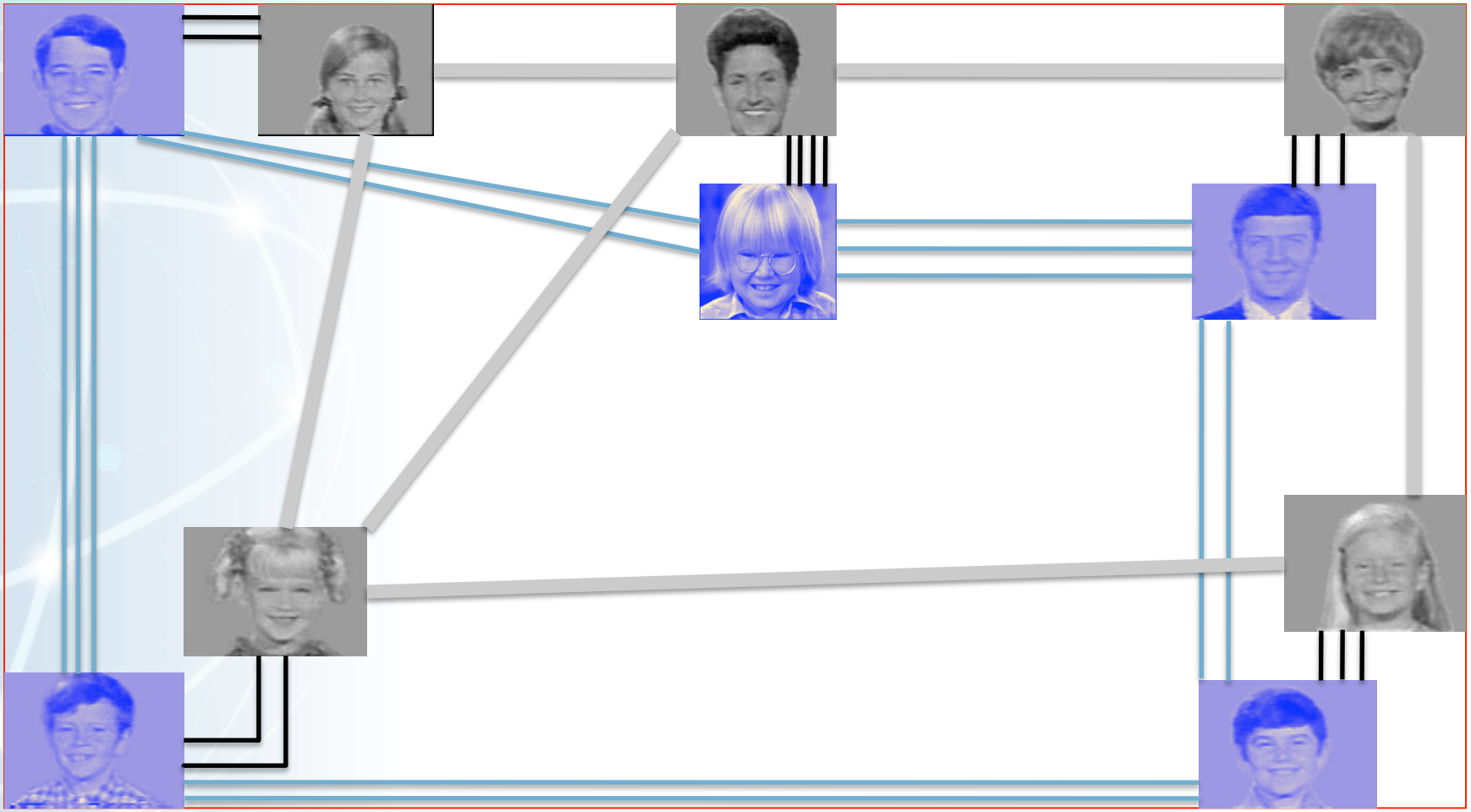




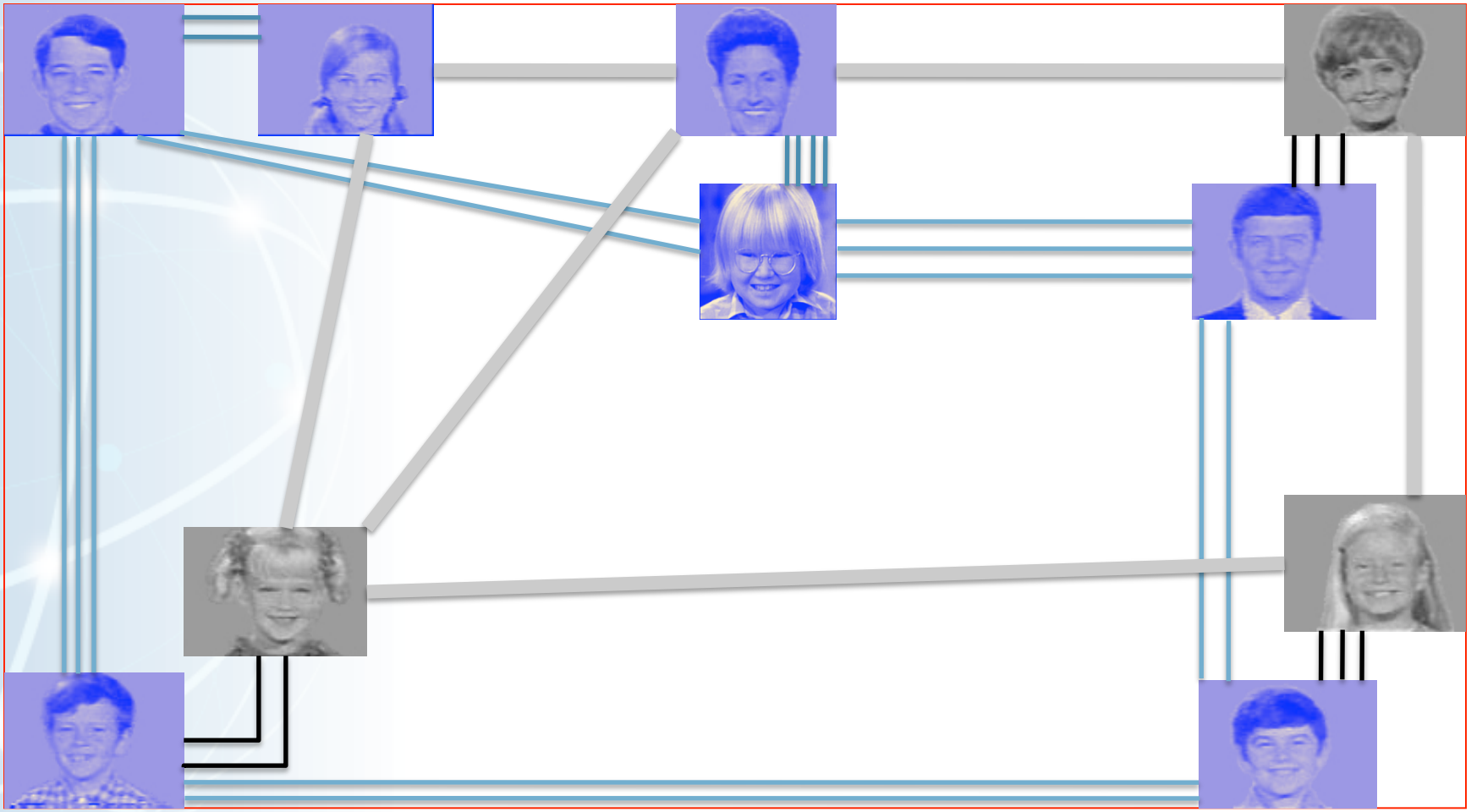
How we formed a network



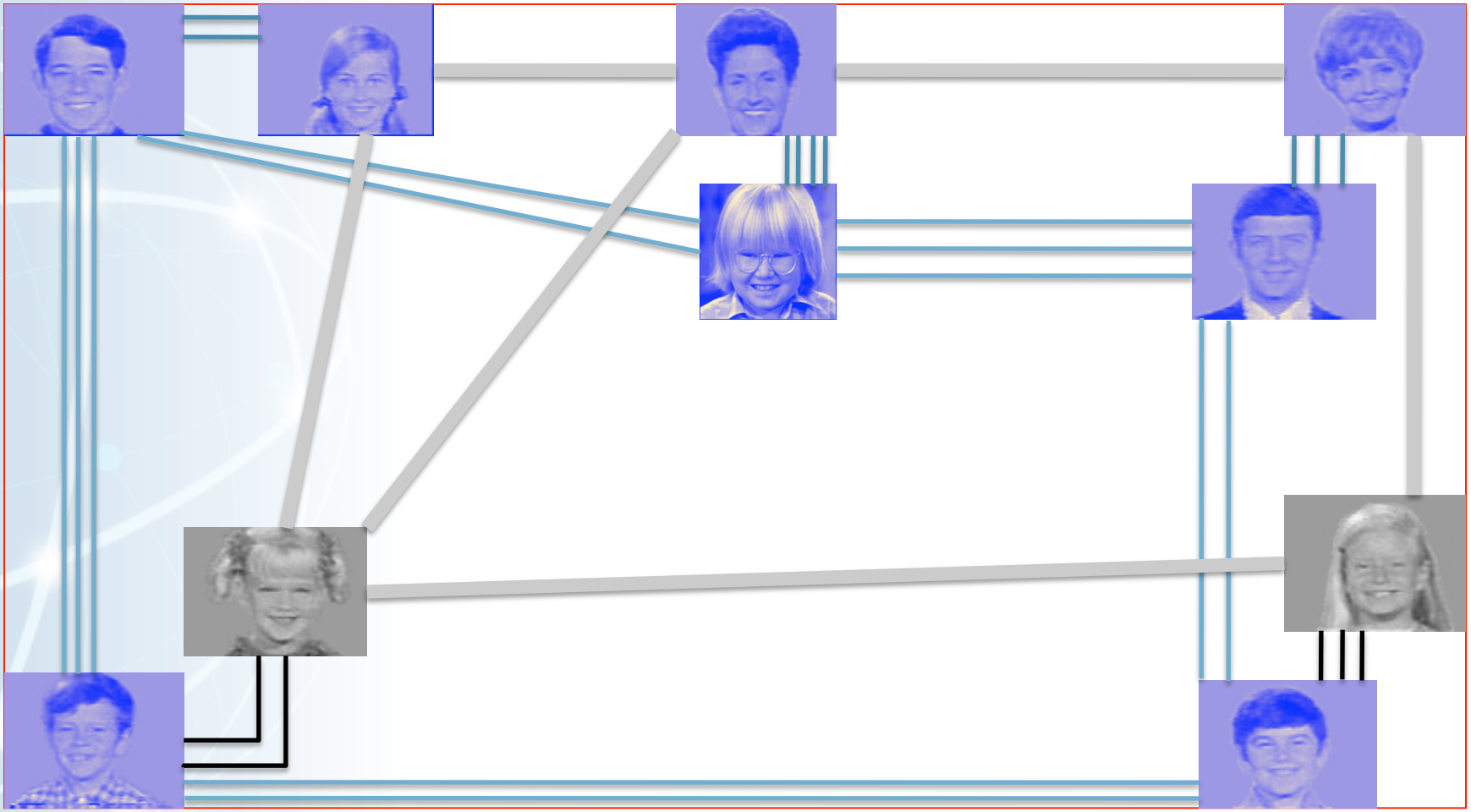
How we formed a network



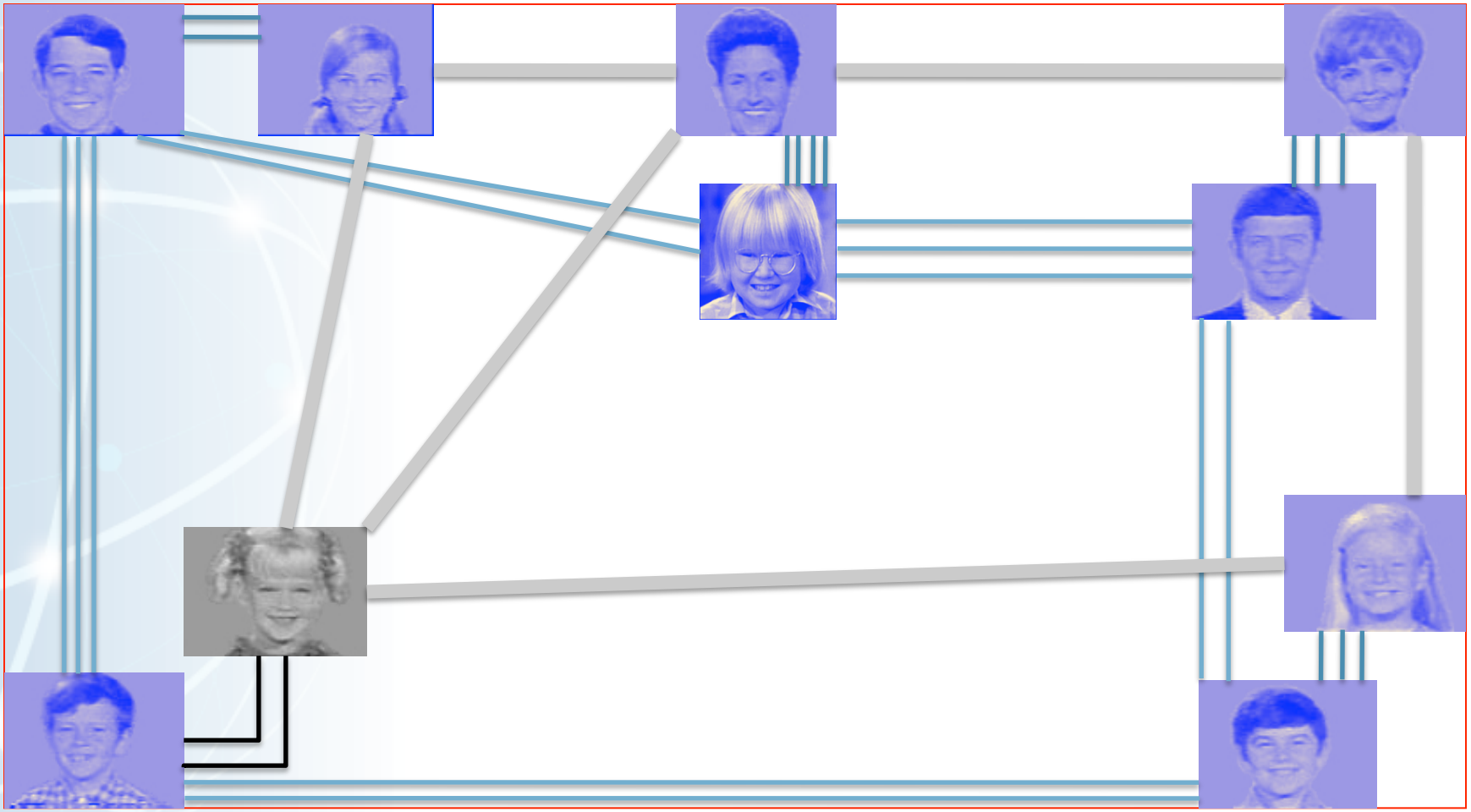
How we formed a network



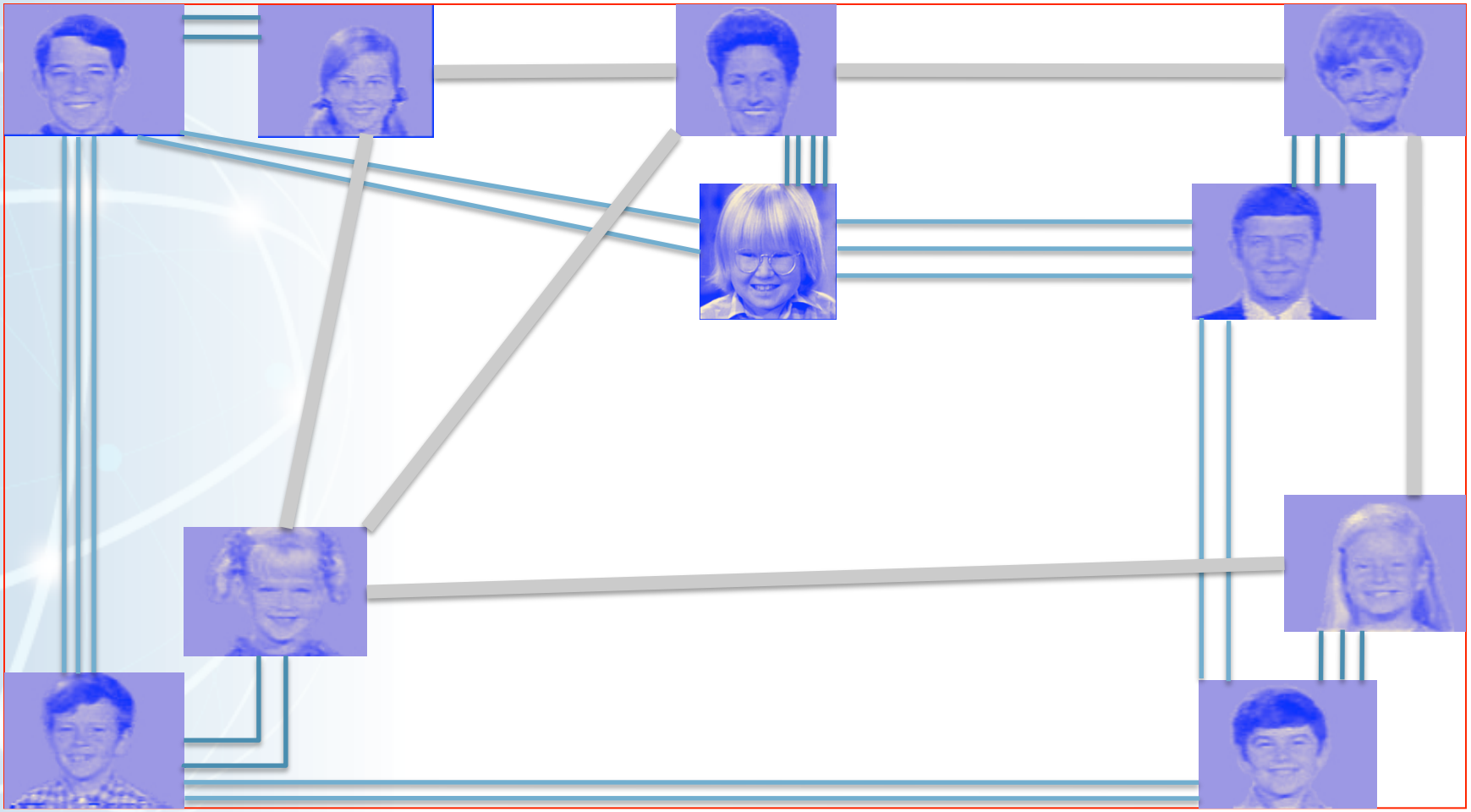
How we formed a network



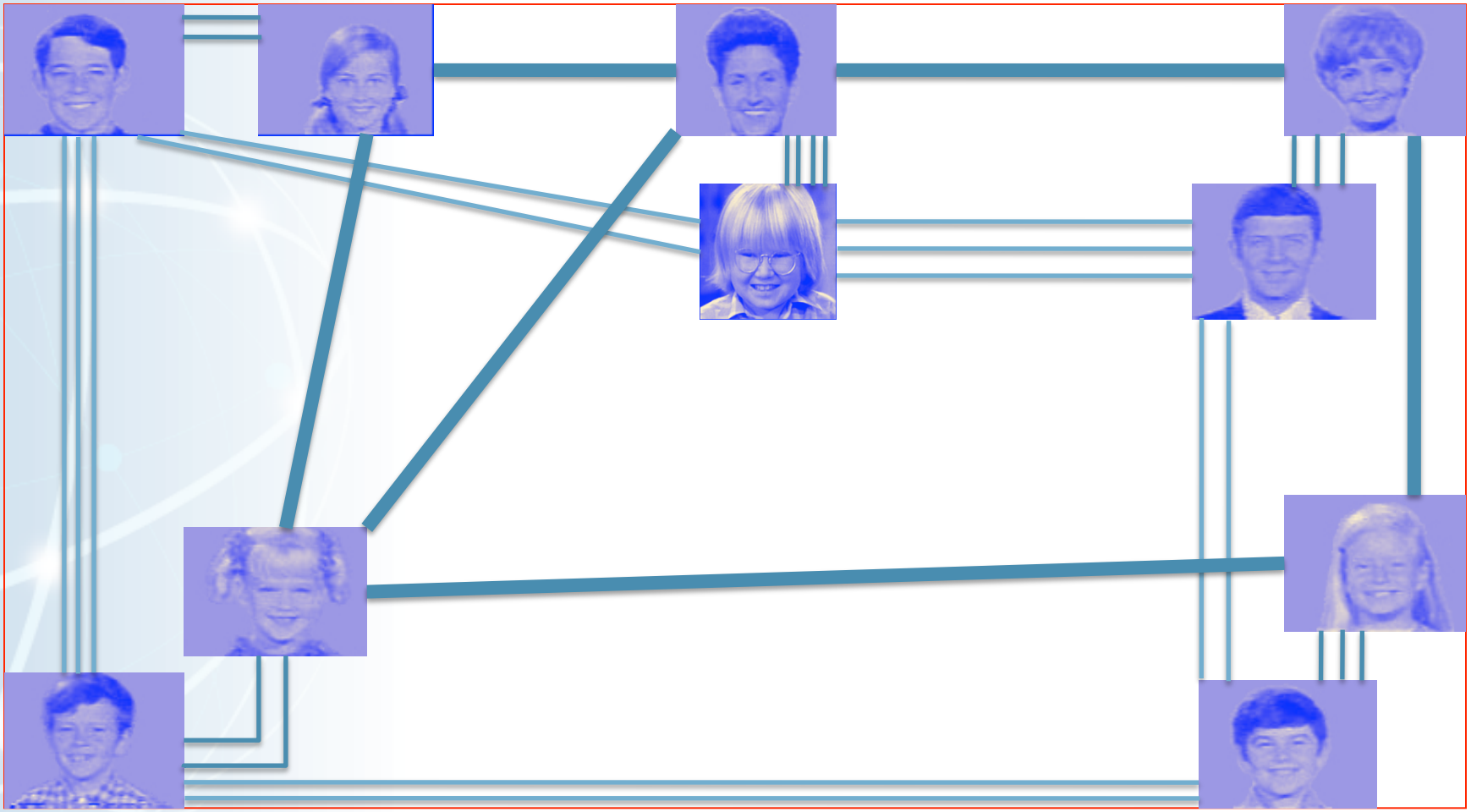
How we formed a network



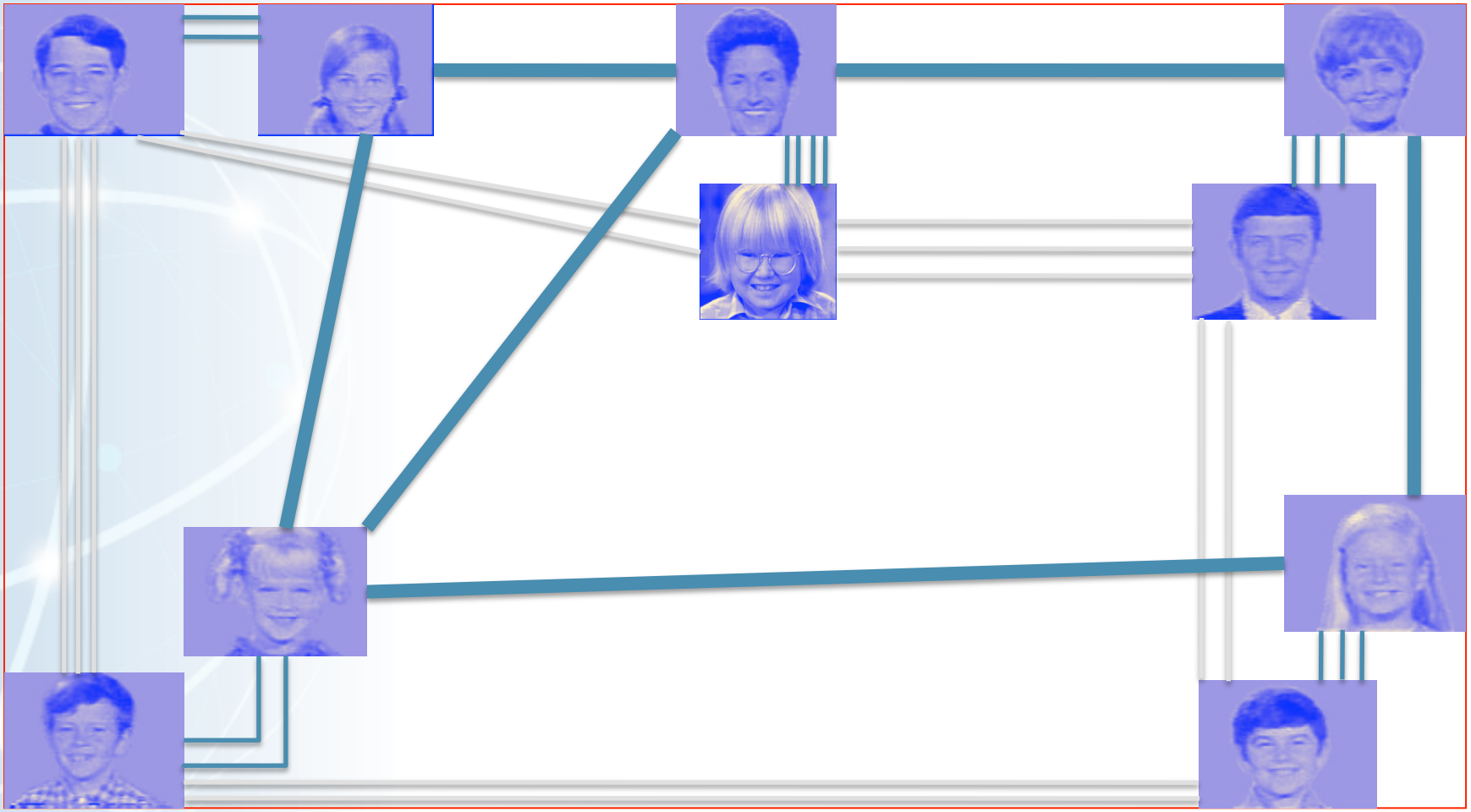
How we formed a network



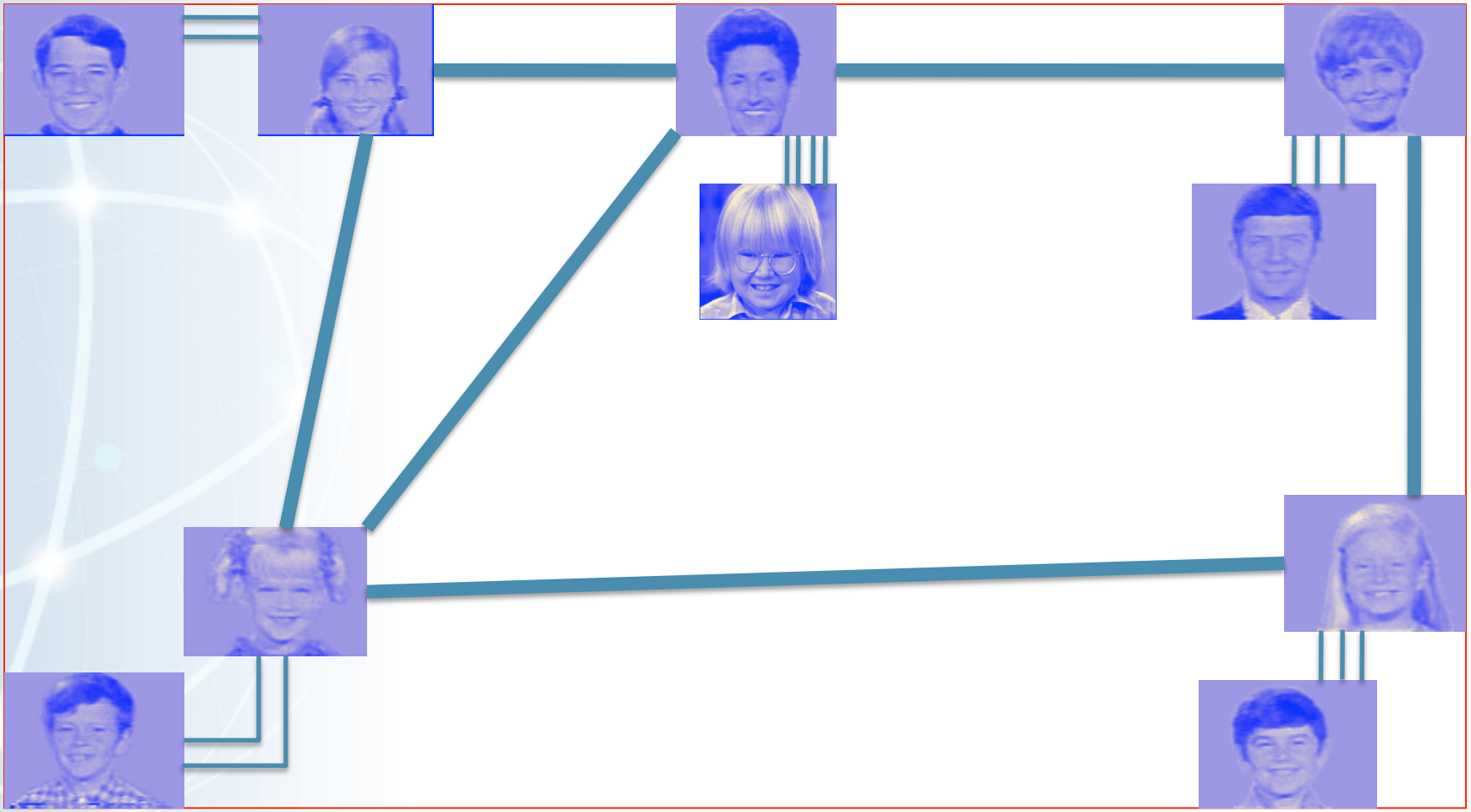
How we formed a network



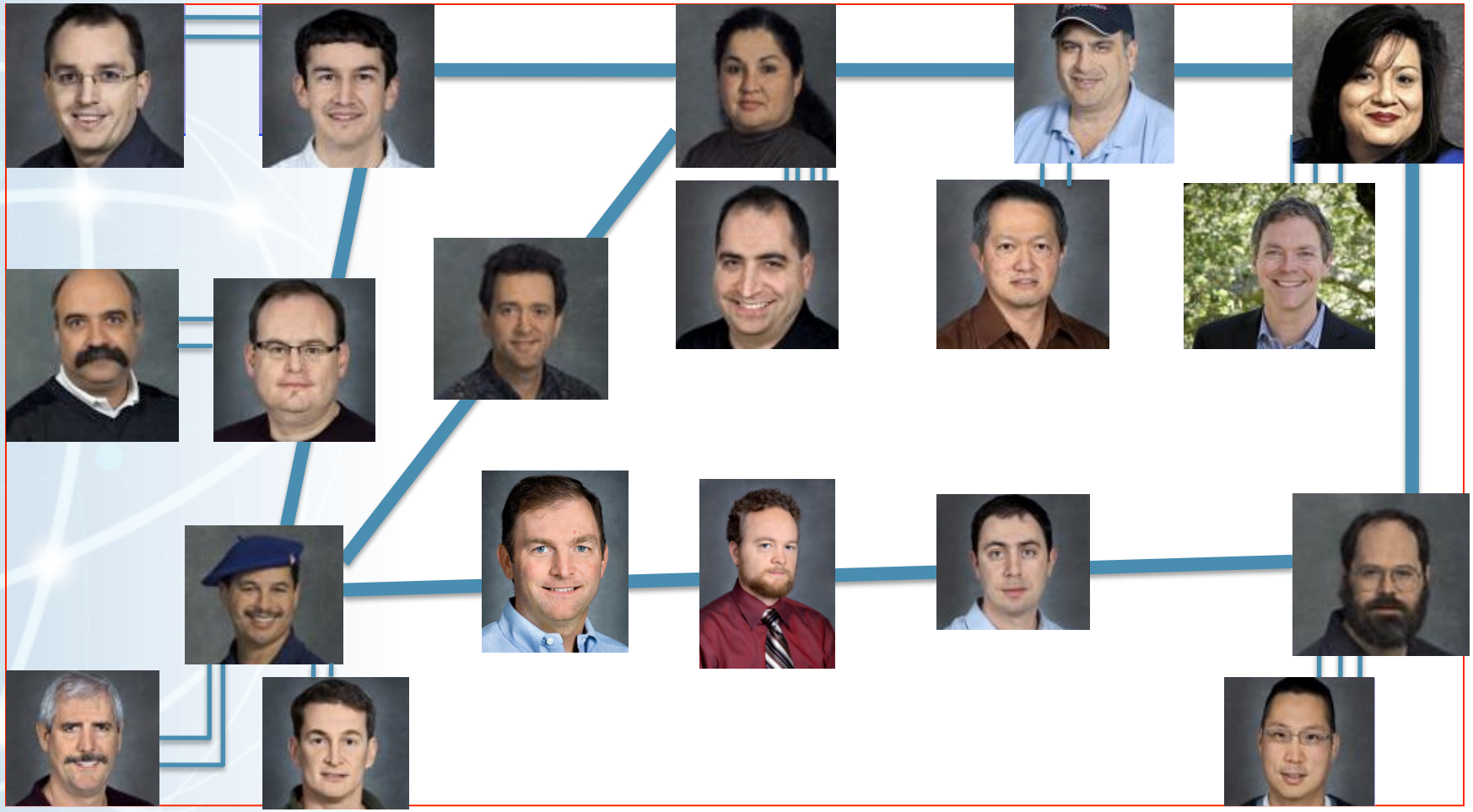
How we formed a network



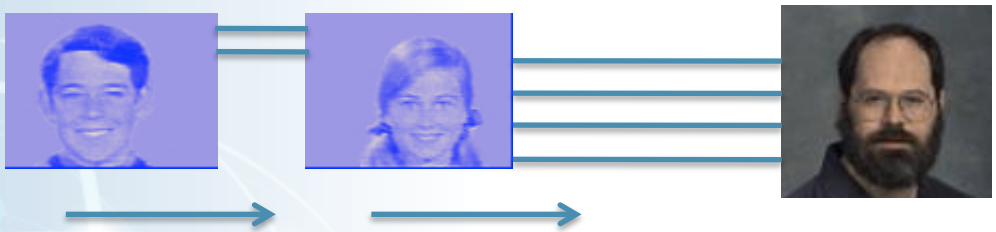
How we formed a network



How we formed a network



Lessons Learned



Routers sometimes have bugs and implementation errors.

- FIB table corruption?

Lessons Learned

Routers sometimes have bugs and implementation errors.

- FIB table corruption?

Lessons Learned

Routers sometimes have bugs and implementation errors.

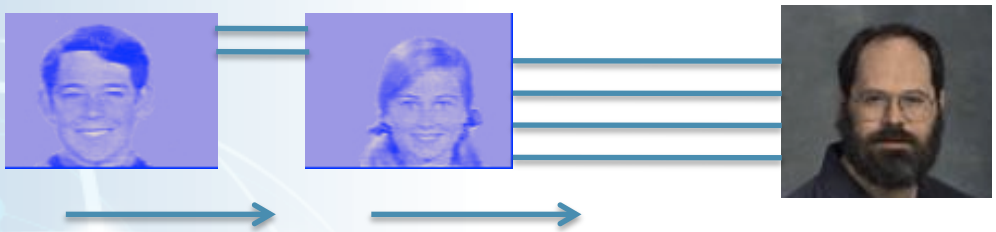
- FIB table corruption?

Lessons Learned

Routers sometimes have bugs and implementation errors.

- FIB table corruption?

Lessons Learned

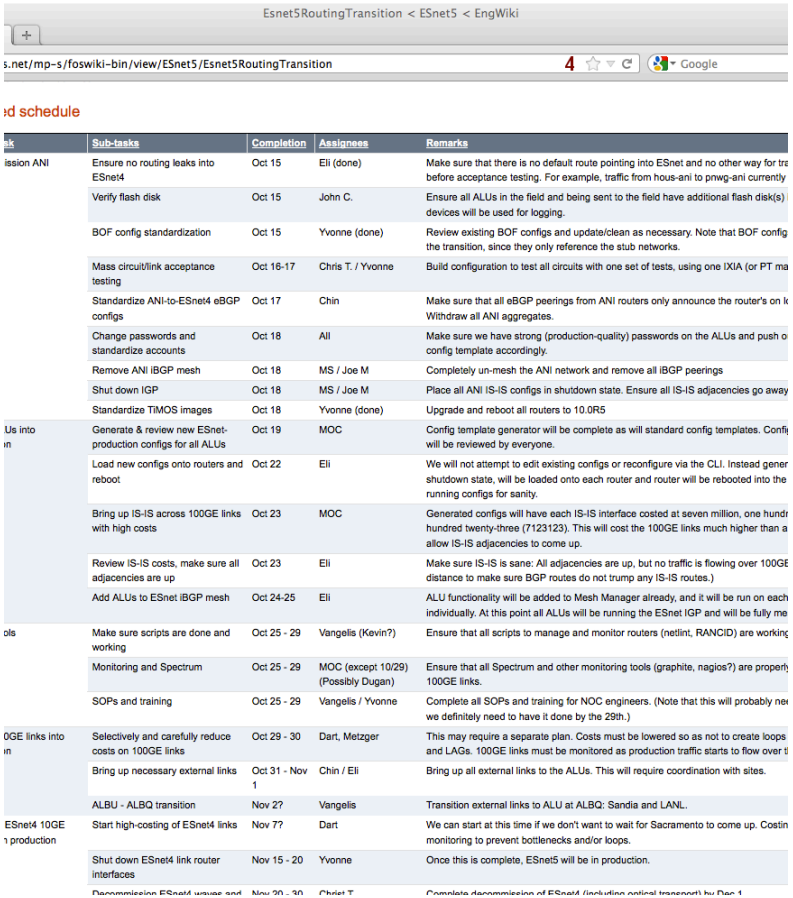


Routers sometimes have bugs and implementation errors.

- FIB table corruption?

Outcomes and Lessons

- More platform varieties (especially old and ignored platforms) leads to complexity and issues. (Shocking!)
- Amazing how well things did work! Shows the benefits of planning.
- Joe Metzger always says “we need a detailed plan!”
- …and he’s absolutely right.


Esnet5RoutingTransition < ESnet5 < EngWiki

s.net/mp-s/foswiki-bin/view/ESnet5/Esnet5RoutingTransition

ed schedule

sk	Sub-tasks	Completion	Assignees	Remarks
ission ANI	Ensure no routing leaks into ESnet4	Oct 15	Eli (done)	Make sure that there is no default route pointing into ESnet and no other way for tra before acceptance testing. For example, traffic from hous-ani to pnwg-ani currently
	Verify flash disk	Oct 15	John C.	Ensure all ALUs in the field and being sent to the field have additional flash disk(s) devices will be used for logging.
	BOF config standardization	Oct 15	Yvonne (done)	Review existing BOF configs and update/clean as necessary. Note that BOF config the transition, since they only reference the stub networks.
	Mass circuit/link acceptance testing	Oct 16-17	Chris T. / Yvonne	Build configuration to test all circuits with one set of tests, using one IXIA (or PT ma
	Standardize ANI-to-ESnet4 eBGP configs	Oct 17	Chin	Make sure that all eBGP peerings from ANI routers only announce the router's on l Withdraw all ANI aggregates.
	Change passwords and standardize accounts	Oct 18	All	Make sure we have strong (production-quality) passwords on the ALUs and push o config template accordingly.
	Remove ANI iBGP mesh	Oct 18	MS / Joe M	Completely un-mesh the ANI network and remove all iBGP peerings
	Shut down IGP	Oct 18	MS / Joe M	Place all ANI IS-IS configs in shutdown state. Ensure all IS-IS adjacencies go away
	Standardize TiMOS images	Oct 18	Yvonne (done)	Upgrade and reboot all routers to 10.0R5
Us into n	Generate & review new ESnet-production configs for all ALUs	Oct 19	MOC	Config template generator will be complete as will standard config templates. Confi will be reviewed by everyone.
	Load new configs onto routers and reboot	Oct 22	Eli	We will not attempt to edit existing configs or reconfigure via the CLI. Instead gener shutdown state, will be loaded onto each router and router will be rebooted into the running configs for sanity.
	Bring up IS-IS across 100GE links with high costs	Oct 23	MOC	Generated configs will have each IS-IS interface costed at seven million, one hundr hundred twenty-three (7123123). This will cost the 100GE links much higher than a allow IS-IS adjacencies to come up.
	Review IS-IS costs, make sure all adjacencies are up	Oct 23	Eli	Make sure IS-IS is sane: All adjacencies are up, but no traffic is flowing over 100GE distance to make sure BGP routes do not trump any IS-IS routes.)
	Add ALUs to ESnet iBGP mesh	Oct 24-25	Eli	ALU functionality will be added to Mesh Manager already, and it will be run on each individually. At this point all ALUs will be running the ESnet IGP and will be fully me
ols	Make sure scripts are done and working	Oct 25 - 29	Vangelis (Kevin?)	Ensure that all scripts to manage and monitor routers (netlint, RANCID) are working
	Monitoring and Spectrum	Oct 25 - 29	MOC (except 10/29) (Possibly Dugan)	Ensure that all Spectrum and other monitoring tools (graphite, nagios?) are properly 100GE links.
	SOPs and training	Oct 25 - 29	Vangelis / Yvonne	Complete all SOPs and training for NOC engineers. (Note that this will probably ne we definitely need to have it done by the 29th.)
0GE links into n	Selectively and carefully reduce costs on 100GE links	Oct 29 - 30	Dart, Metzger	This may require a separate plan. Costs must be lowered so as not to create loops and LAGs. 100GE links must be monitored as production traffic starts to flow over t
	Bring up necessary external links	Oct 31 - Nov 1	Chin / Eli	Bring up all external links to the ALUs. This will require coordination with sites.
	ALBU - ALBQ transition	Nov 2?	Vangelis	Transition external links to ALU at ALBQ: Sandia and LANL.
ESnet4 10GE n production	Start high-costing of ESnet4 links	Nov 7?	Dart	We can start at this time if we don't want to wait for Sacramento to come up. Costin monitoring to prevent bottlenecks and/or loops.
	Shut down ESnet4 link router interfaces	Nov 15 - 20	Yvonne	Once this is complete, ESnet5 will be in production.
	Decommission ESnet4 waves and	Nov 20 - 30	Christ T	Complete decommission of ESnet4 (including optical transport) by Dec 1

Thanks!

Michael Sinatra

http://www.es.net/

http://fasterdata.es.net/