iSCSI SANs Don’t Have To Suck

Derek J. Balling
Data Center Manager
derekb@answers.com

Answers.com®

What is iSCSI?

- iSCSI is a network-based block-level disk protocol
- Essentially SCSI commands stuffed into the payload of TCP packets

When Can iSCSI Typically Suck?

- iSCSI is extremely vulnerable to latency and even super-short (millisecond) interruptions, just as conventional SCSI disks might be problematic if the cable between the controller and disks didn't have 100% reliability
- Ethernet networks often have bursts of poor performance (latency) and interruptions
- Principally, network issues are the main cause of iSCSI pain and suffering

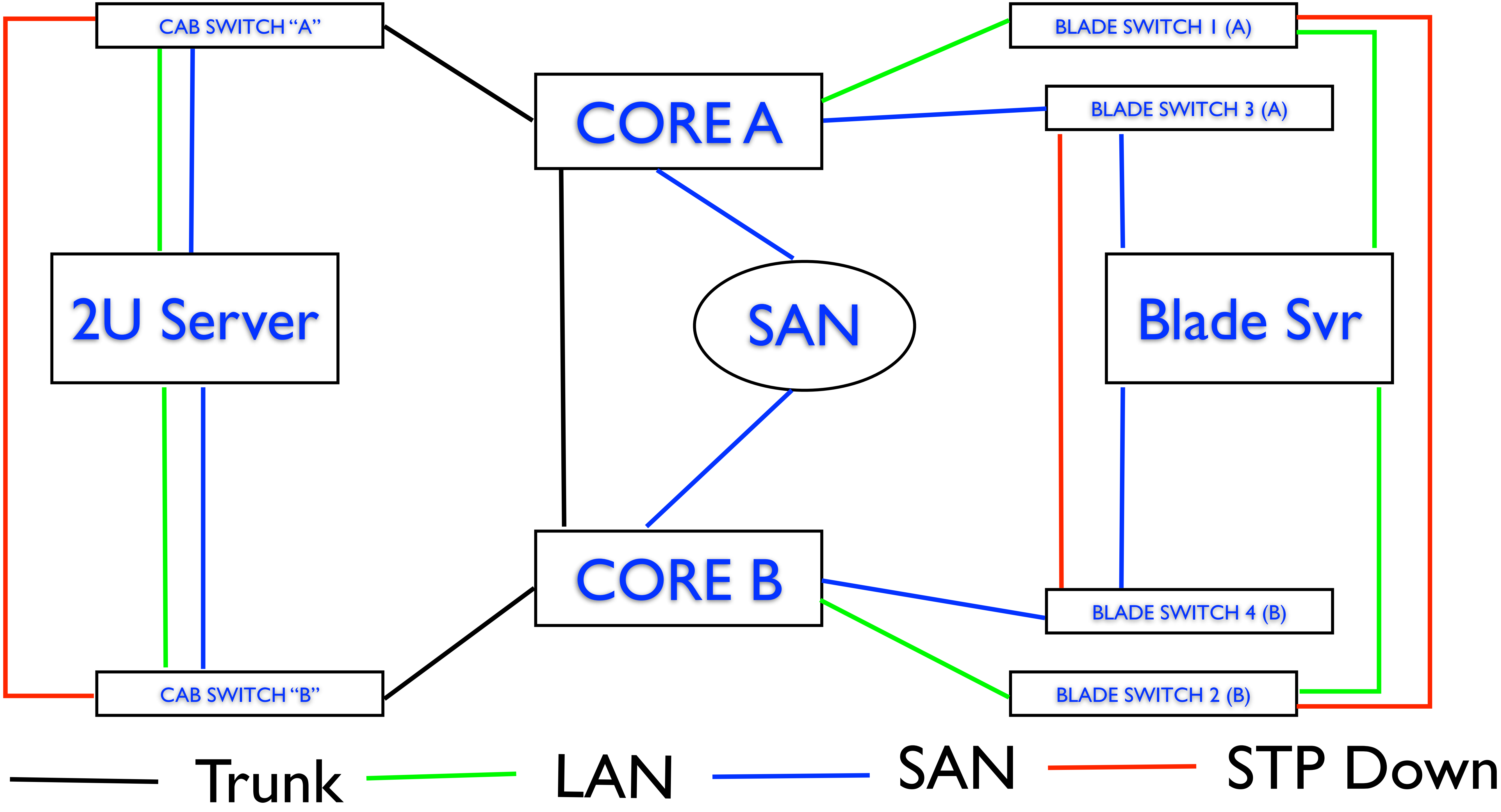
How To Make iSCSI Not Suck

- Need to build a network infrastructure with near-zero outage or packet-loss.
- Great for iSCSI SANs, but the same principles apply for any normal data LAN.
- Really could have called this talk "How To Build A Really Robust Ethernet Network", but it just doesn't capture the level of effect this has on iSCSI
- This is all stuff you already know, but may not have actually put it all together

Our Server Design Principles

- Every machine has four NICs, two “data-network” and, if it needs access to the SAN, two for the “SAN”
- Each network has “A” and “B” sides for redundancy
- “A” and “B” side NICs are in a bonded-pair using active/passive failover

The Initial Design



CAB SWITCH “A”

CORE A

BLADE SWITCH 1 (A)

BLADE SWITCH 3 (A)

2U Server

SAN

Blade Svr

CORE B

BLADE SWITCH 4 (B)

CAB SWITCH “B”

BLADE SWITCH 2 (B)

Trunk LAN SAN STP Down

The Initial Network Design

- Common “Core” switching gear between data-network and SAN
- Multiple VLANs, mostly on the data-network side, but one VLAN for the SAN traffic
- Dual Cabinet Switches / Quad Blade Switches

Some Things To Note

- Each NIC in a Blade maps to an individual switch in the enclosure, so there are two "A" side switches and two "B" side switches. The only difference is which port VLANs are mapped to (data-network or SAN)
- The SAN appliances are directly connected to the Core switches
- The links connecting Cab-A/Cab-B, BladeSW1/ BladeSW2, and BladeSW3/BladeSW4 are "inactive" via Spanning Tree

What is Spanning Tree Protocol?

- At the macro level, it’s a protocol that switches use to communicate with each other to ensure that there are no “loops” in the switching fabric
- Where a “loop” exists, it figures out which links to disable to make the loop go away
- Can be configured to prioritize certain links over other links
- Internally we refer to it as controlling the links which “cross the A/B divide” since that’s what causes the actual loop.

Benefits of This Architecture

- Every device has multiple, redundant paths to everything it needs
- Spanning-Tree Protocol ensures that "low-priority" (failover) links stay down until they are needed

Problems We Noticed

- Only one really.
- Spanning Tree Protocol

The Problem: Spanning Tree Events

- Every time a switch is connected, and most times a switch is removed, every switch on the fabric does a quick re-evaluation of what the network looks like
- Generally speaking they don’t pass packets while they’re doing this, other than their own STP packets
- iSCSI is moderately unsuccessful at staying up while the switches refuse to send its packets

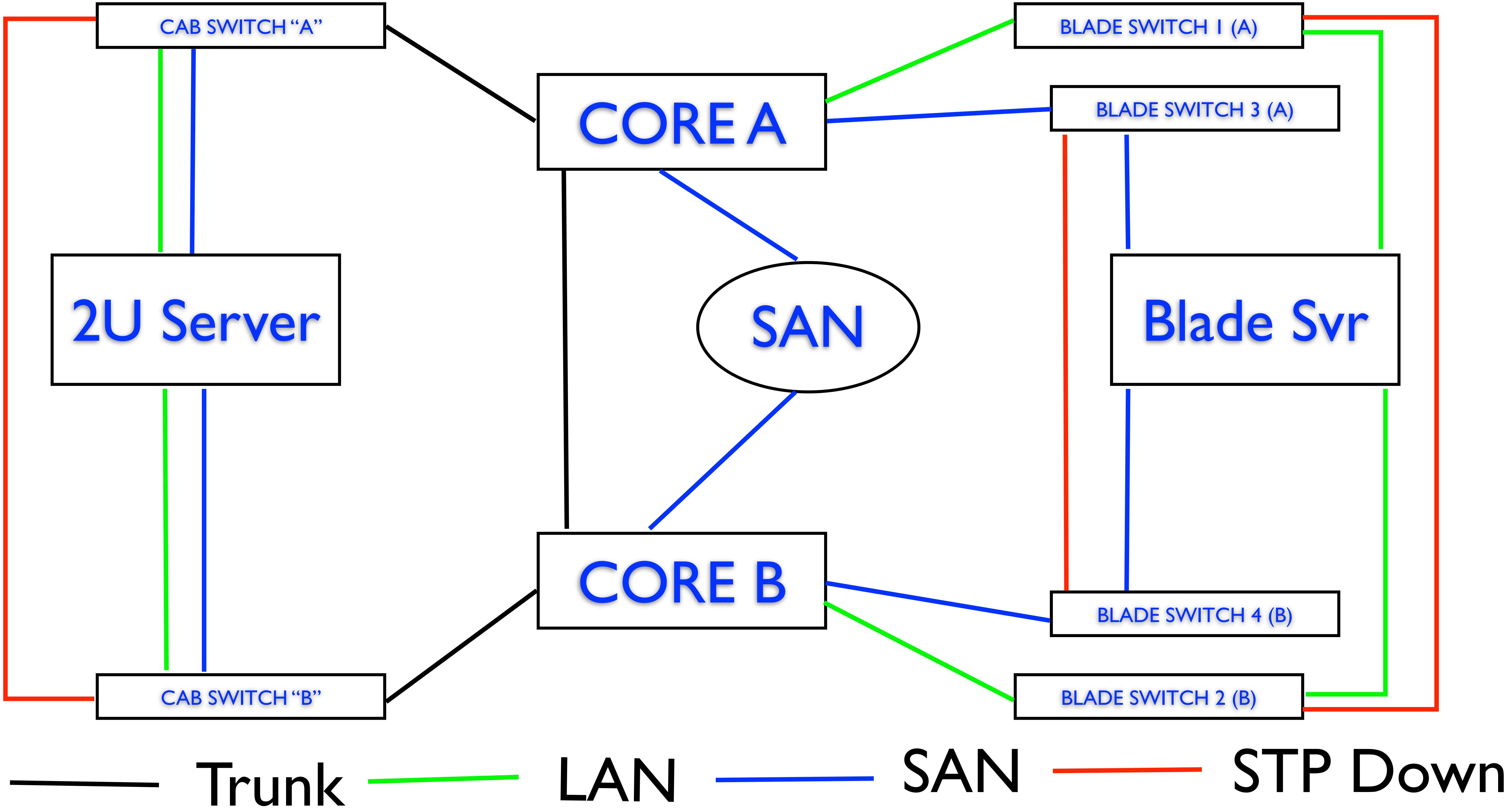
Low Hanging Fruit

- Biggest cause of STP for us was new blade chassis being installed during roll-out
- For Blade Switches, disabling Spanning Tree Protocol and enabling instead "Uplink Failure Detection"
- Instead of having the "A" side switch hand traffic over to the "B" side switch to get up to the cores, let the servers just immediately notice the outage and direct traffic directly to the "B" side network equipment

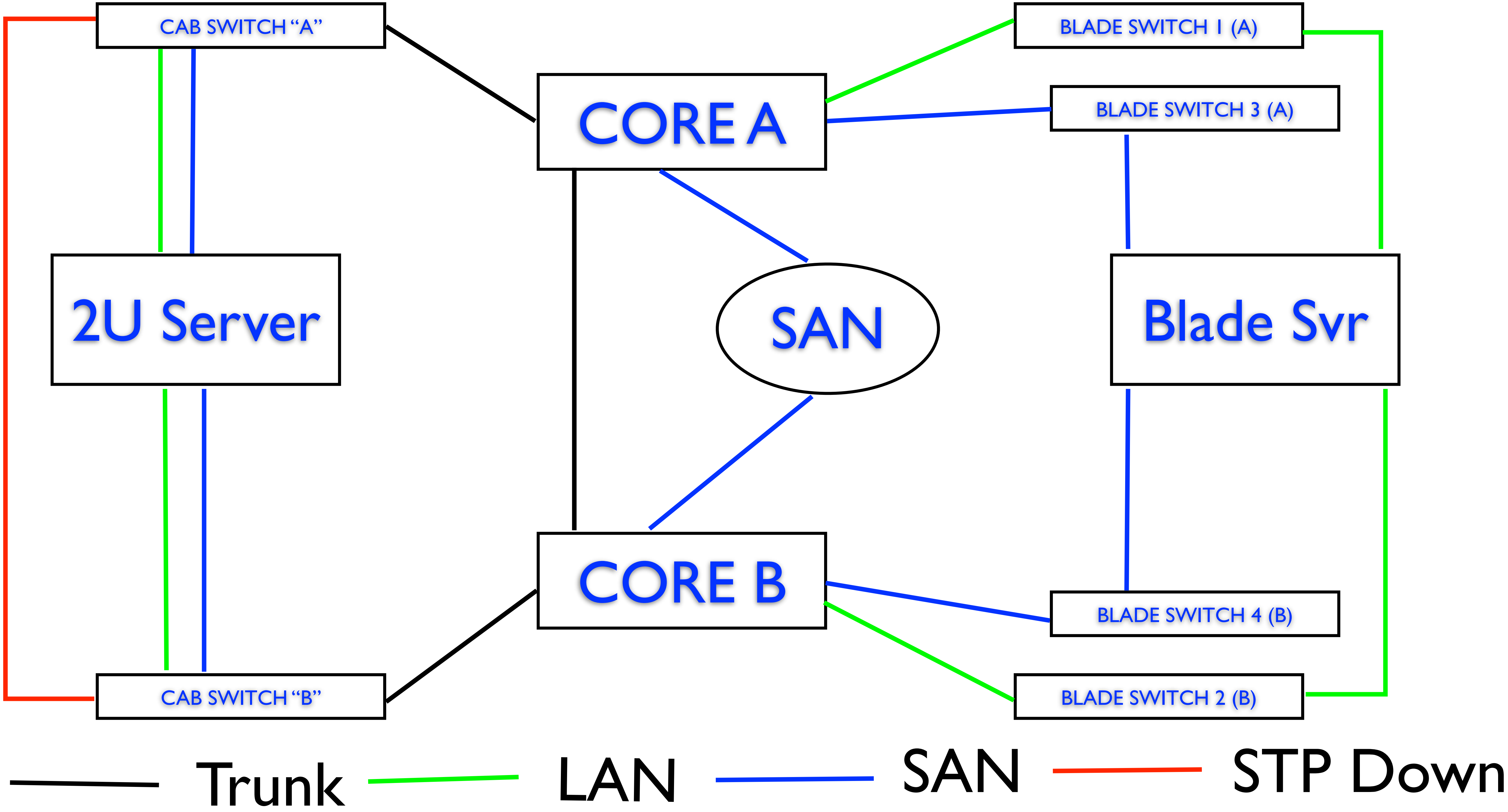
Uplink Failure Detection

- Feature of the Blade Networks blade switches. Juniper and Cisco appear to also support it on some of their product line.
- Switch has two categories of ports, "Link To Monitor" (LTM) and "Link To Disable" (LTD)
- If the "Link" on the LTM ports (or a LACP group) goes dark, it immediately disables all ports in the LTD group
- Put the core-uplink port-channel in the LTM group, the blades in the LTD group

Before Uplink Failure Detection



After Uplink Failure Detection



After Uplink Failure Detection

- Lots of STP events went away, since the Blade Switches no longer "participated" in the STP negotiation
- Connecting new blade chassis to the network didn't trigger an STP "event", meaning iSCSI didn't see as many problems
- Still not 100% success - we still need to install cabinet switches from time to time, and they don't have Uplink Failure Detection, and any network maintenance is extremely problematic

The Ultimate Decision

- We want/need spanning tree on our data LAN so that our servers in standard "pizza-box" cabinets can have redundant upstream links, without all needing to be consuming expensive core switchports
- We don't want it on the SAN, at all
- We're almost never using our 2U servers as SAN consumers
- Build out a new, flat, network, for the SAN. For the few 2Us that need to connect to it, we'll jack them into the new "SAN Cores"

The Plan to Eliminate STP

- The dreaded phrase, “Flat Network”
- Done right, and within certain scales, it can work just fine
- Lots of network folks will tell you, it’s bad, it’s wrong, etc., but it seems to have been the right solution for us

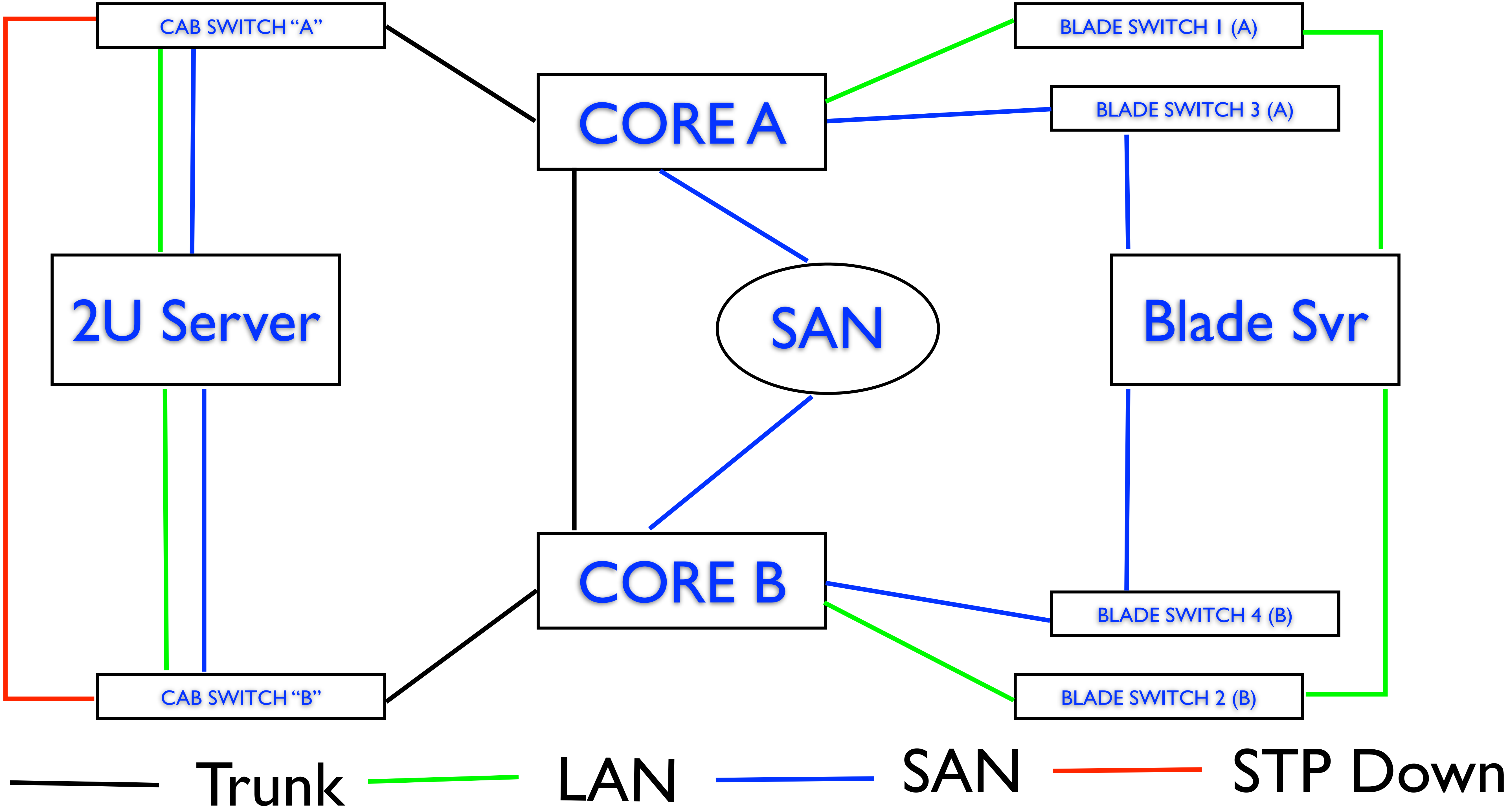
What It Will Look Like

- Small number of 2U Consumers directly connected to the “A” and “B” side “SAN Core” switches
- “A” and “B” side SAN Core switches interconnected
- “A” and “B” side SAN Blade switches connected only to their consumer blades and to their respective core
- Only one “A/B Bridge” - No loops, no STP needed

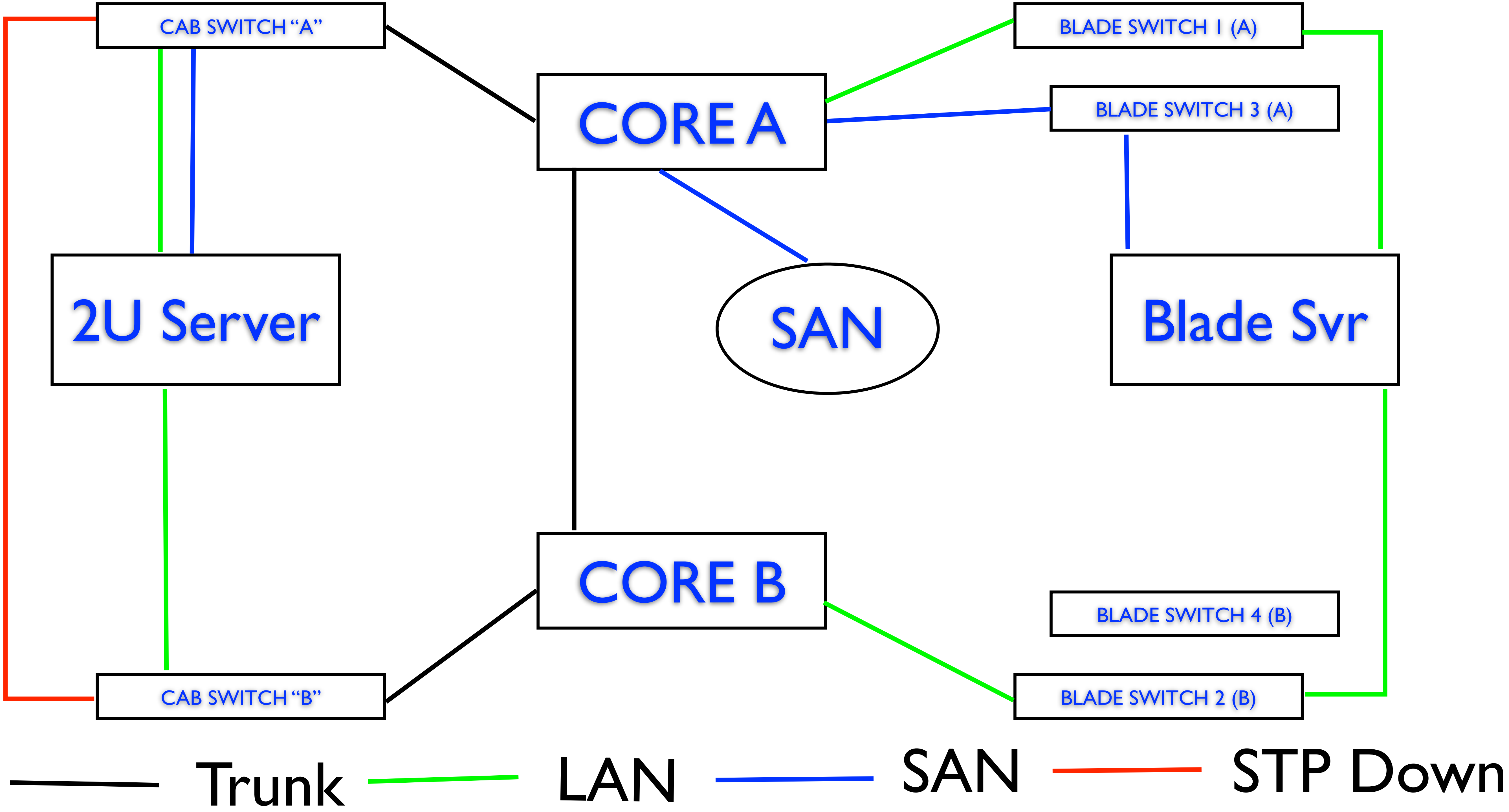
How Do We Get There?

- This is where it gets a little tricky to visualize
- We can disable and isolate any given piece of hardware in our network environment safely
- Once a piece of hardware has been isolated, we can swap it out for new hardware
- “Swap” here can also simply mean “move the cables to some other similarly isolated new piece of hardware”

Step By Step Walk-Through



Disable All SAN “B” Sides and Disconnect



CAB SWITCH “A”

CORE A

BLADE SWITCH 1 (A)

BLADE SWITCH 3 (A)

2U Server

SAN

Blade Svr

CORE B

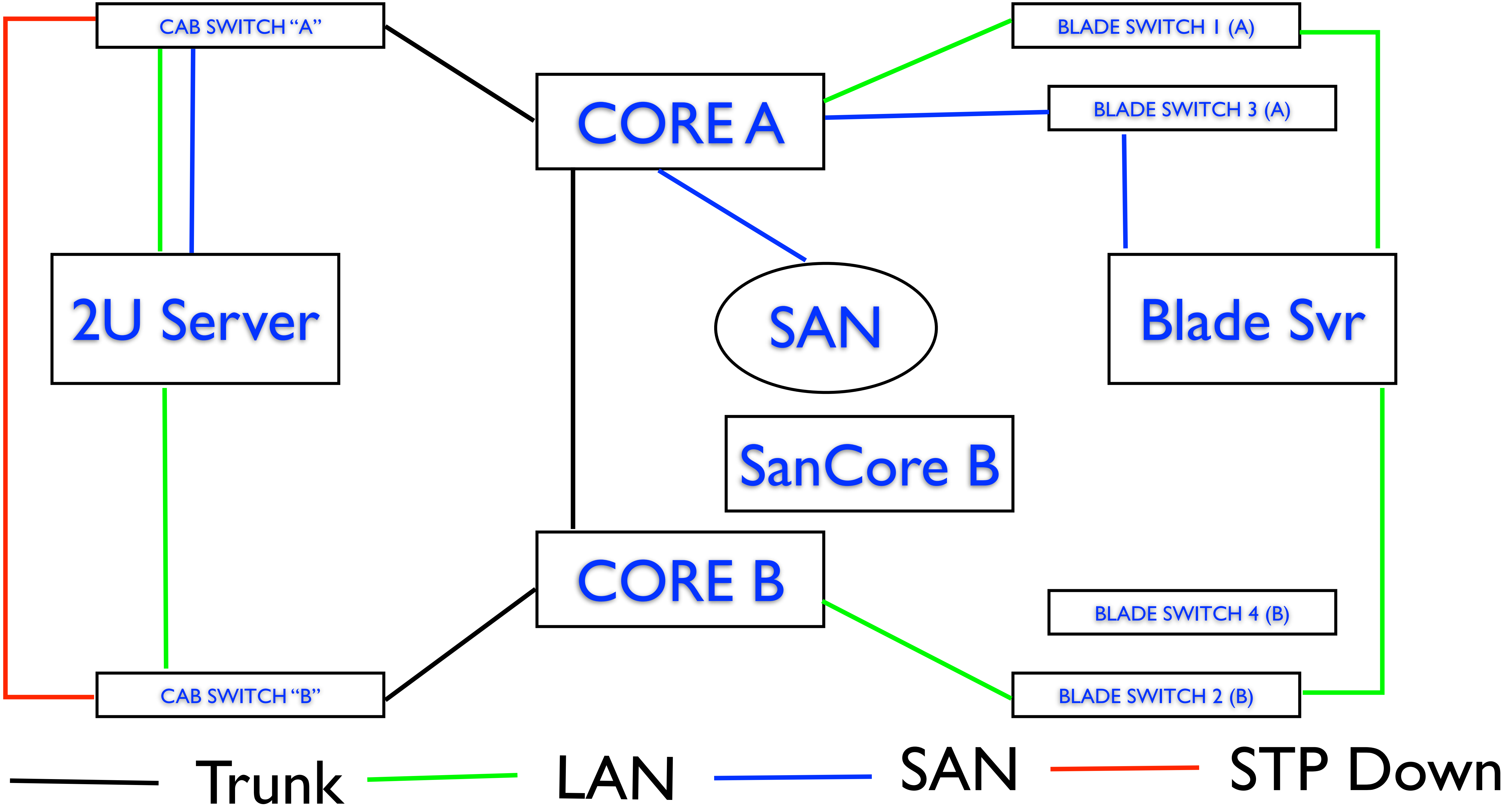
BLADE SWITCH 4 (B)

CAB SWITCH “B”

BLADE SWITCH 2 (B)

Trunk LAN SAN STP Down

Install New "B" Side "SANCore" Switch



CAB SWITCH "A"

CORE A

BLADE SWITCH 1 (A)

BLADE SWITCH 3 (A)

2U Server

SAN

Blade Svr

SanCore B

CORE B

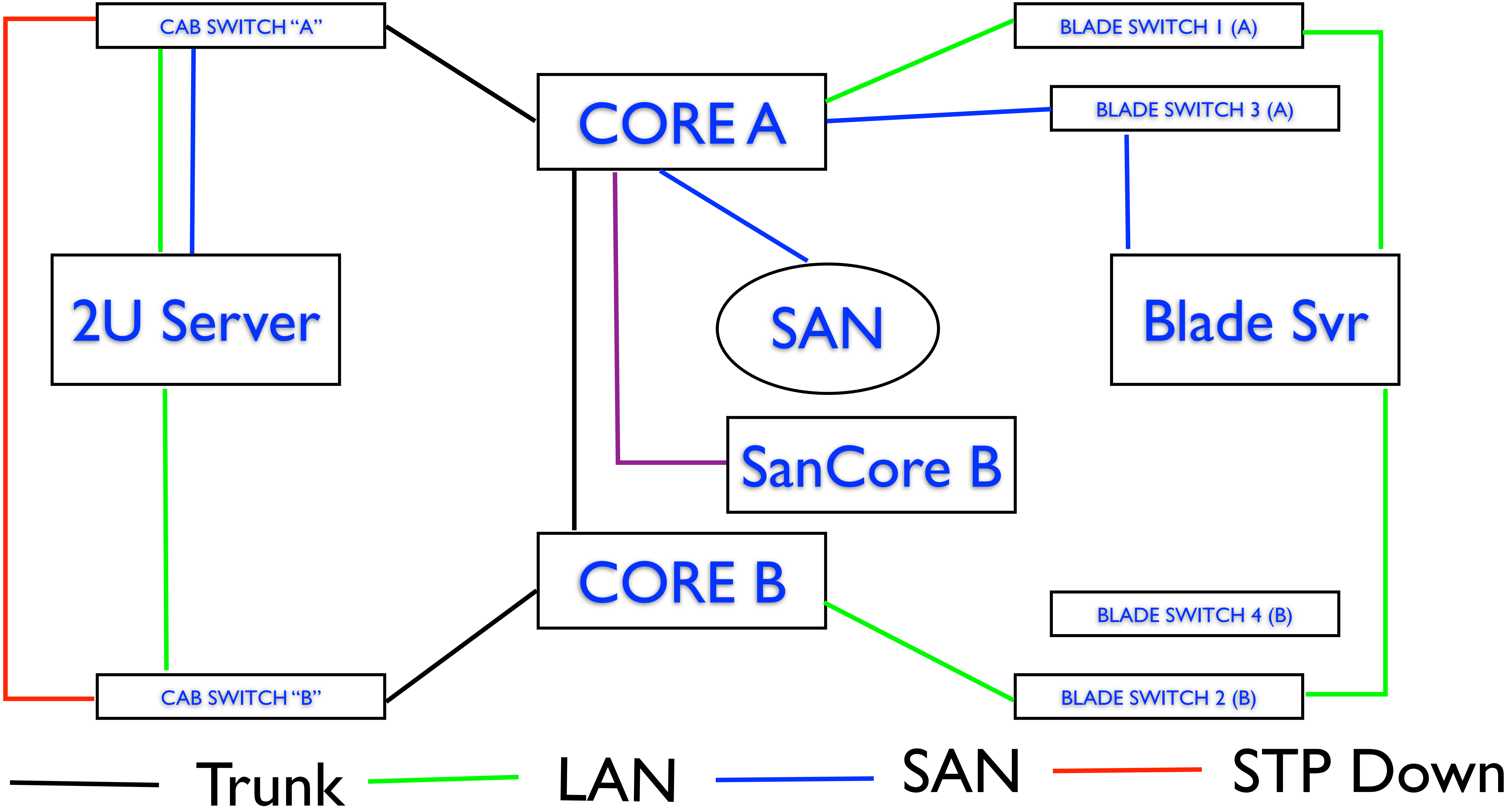
BLADE SWITCH 4 (B)

CAB SWITCH "B"

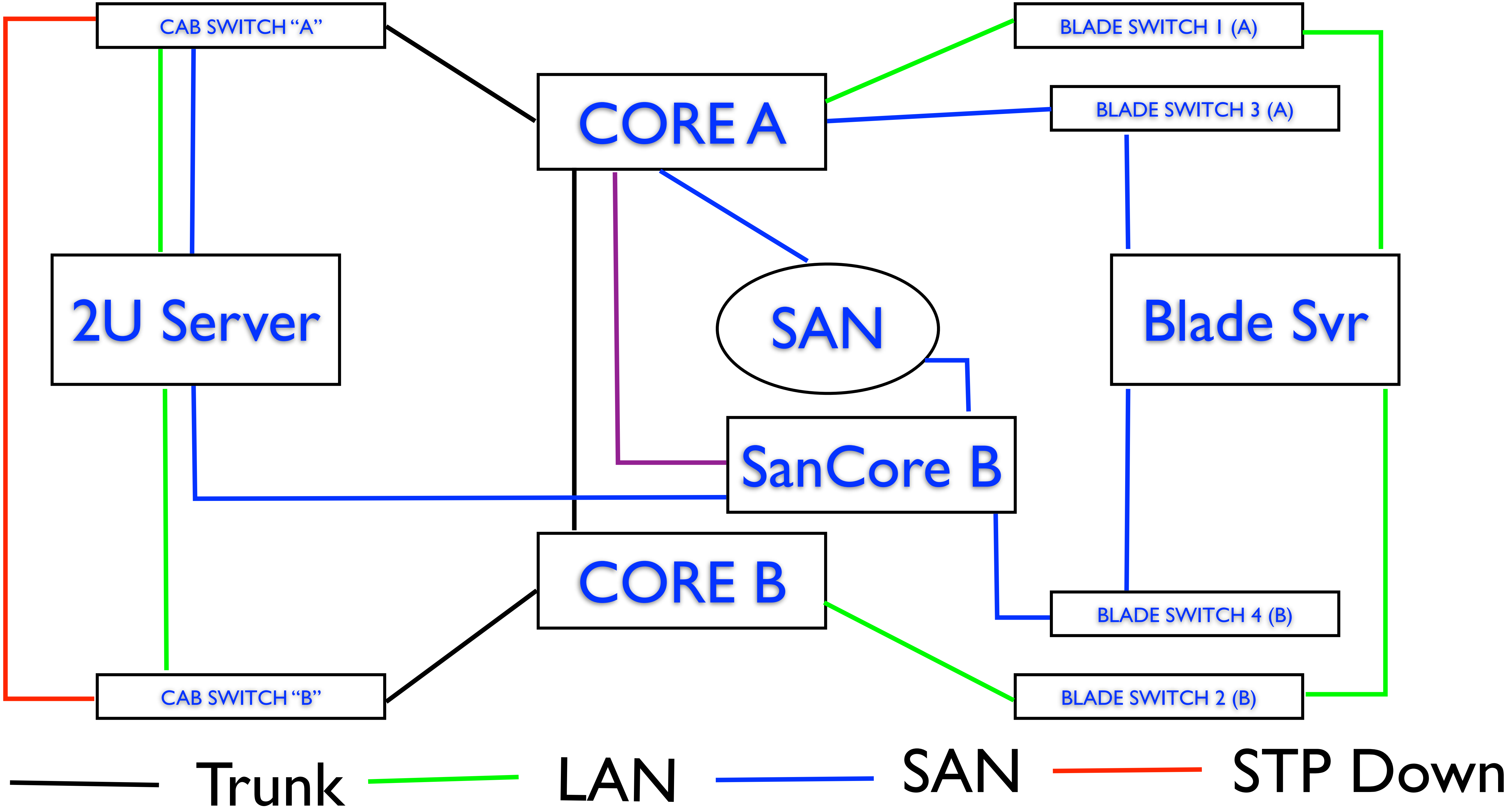
BLADE SWITCH 2 (B)

⬛ Trunk 🟩 LAN 🟦 SAN 🟥 STP Down

Connect Temp Cable From “A” Core to “B” SanCore



Connect “B” Side SAN Equipment to SanCore B



Step-By-Step Example

- Disable all the “B” side SAN links on the 2U and blade consumers, as well as the SAN modules themselves
- Install the new “B” side “SANCore” Switch near the existing “B” side Core switch
- KEY! Connect a temporary cable from the “A” side “Core” to the “B” side “SANCore”
- Move all the “B” side SAN cables from the “B” side “Core” to the “B” side “SANCore”.

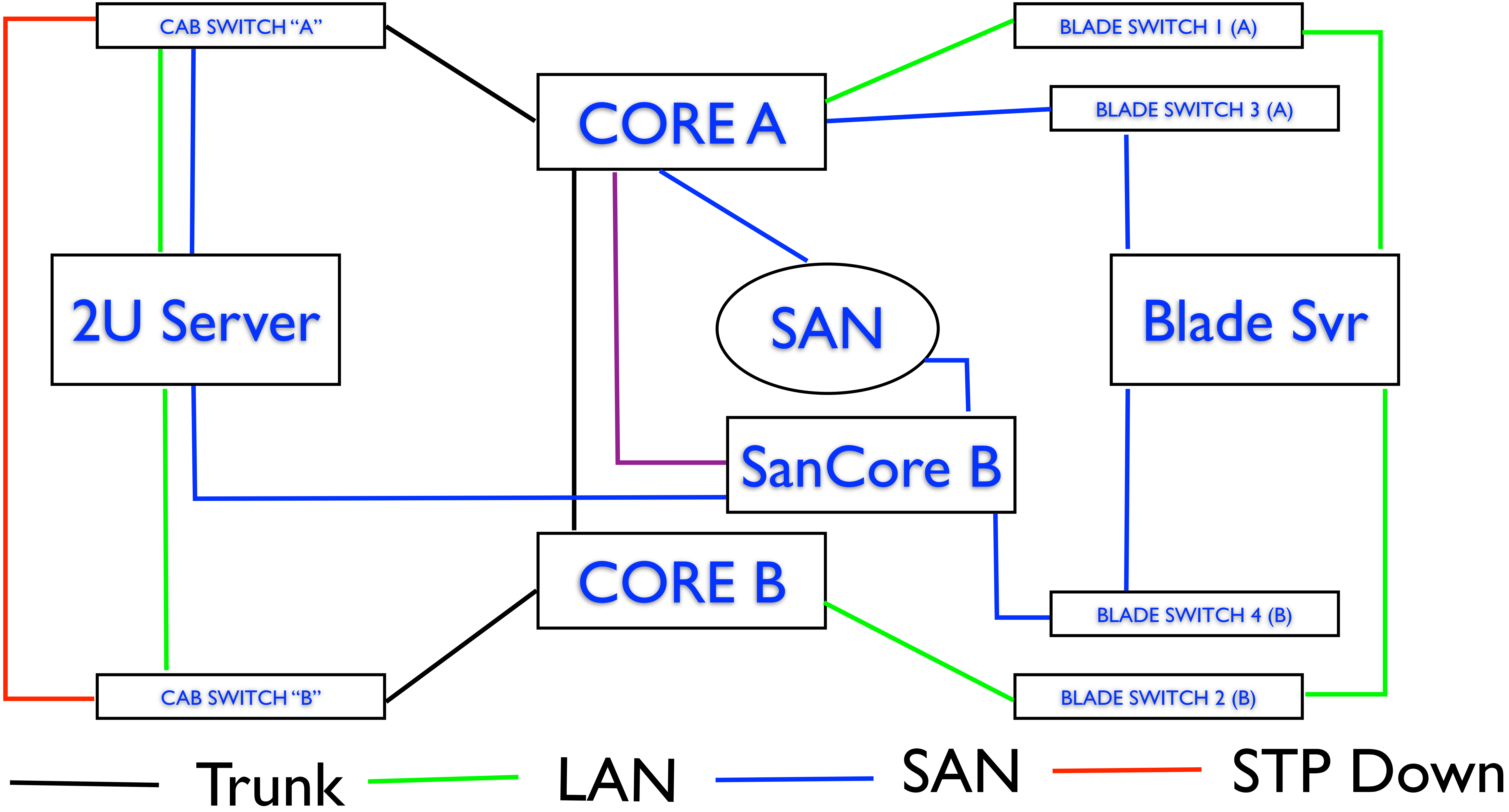
Why The Temporary Cable?

- You're working in a load-balanced/NIC-teaming environment
- Packets might originate on the "A" side for MAC addresses that are presenting themselves on the "B" side hardware
- You definitely don't want any piece of hardware to have its "A" and "B" side NICs on networks that can't see each other, especially when your systems all expect that they can do so.

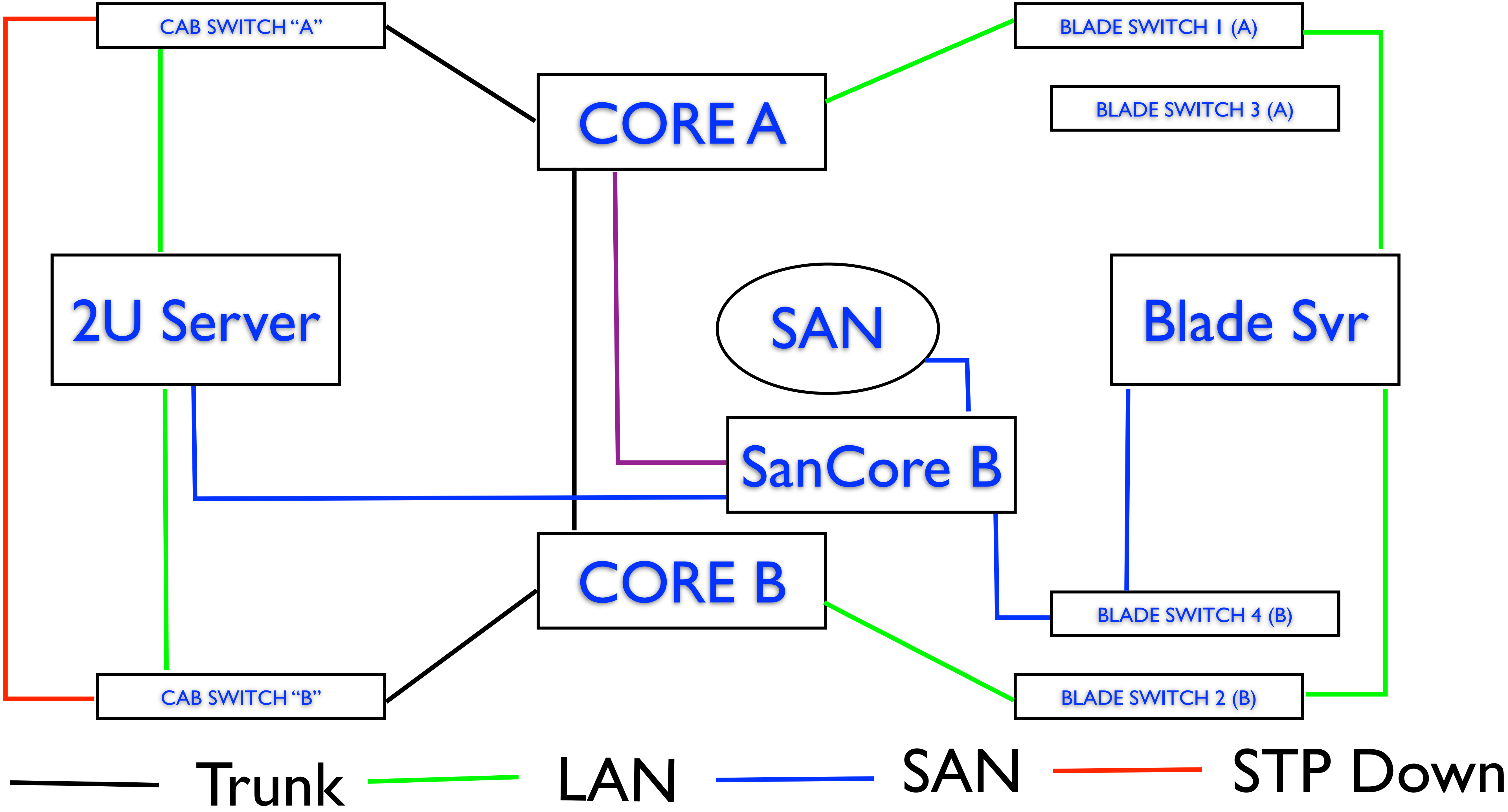
What You've Got Right Now

- Right now you've got this hybrid FrankenNetwork, with "B" side NICs connected to their own 'independent' network
- Light up all the "B" side NICs, ports, etc. Run for a little while on this hybrid network and let things settle down
- But, that temporary cable is your lifeblood right now, because you don't want to *separate* live "A" and "B" networks ever. Badness and pain will ensue
- Lather, rinse, repeat

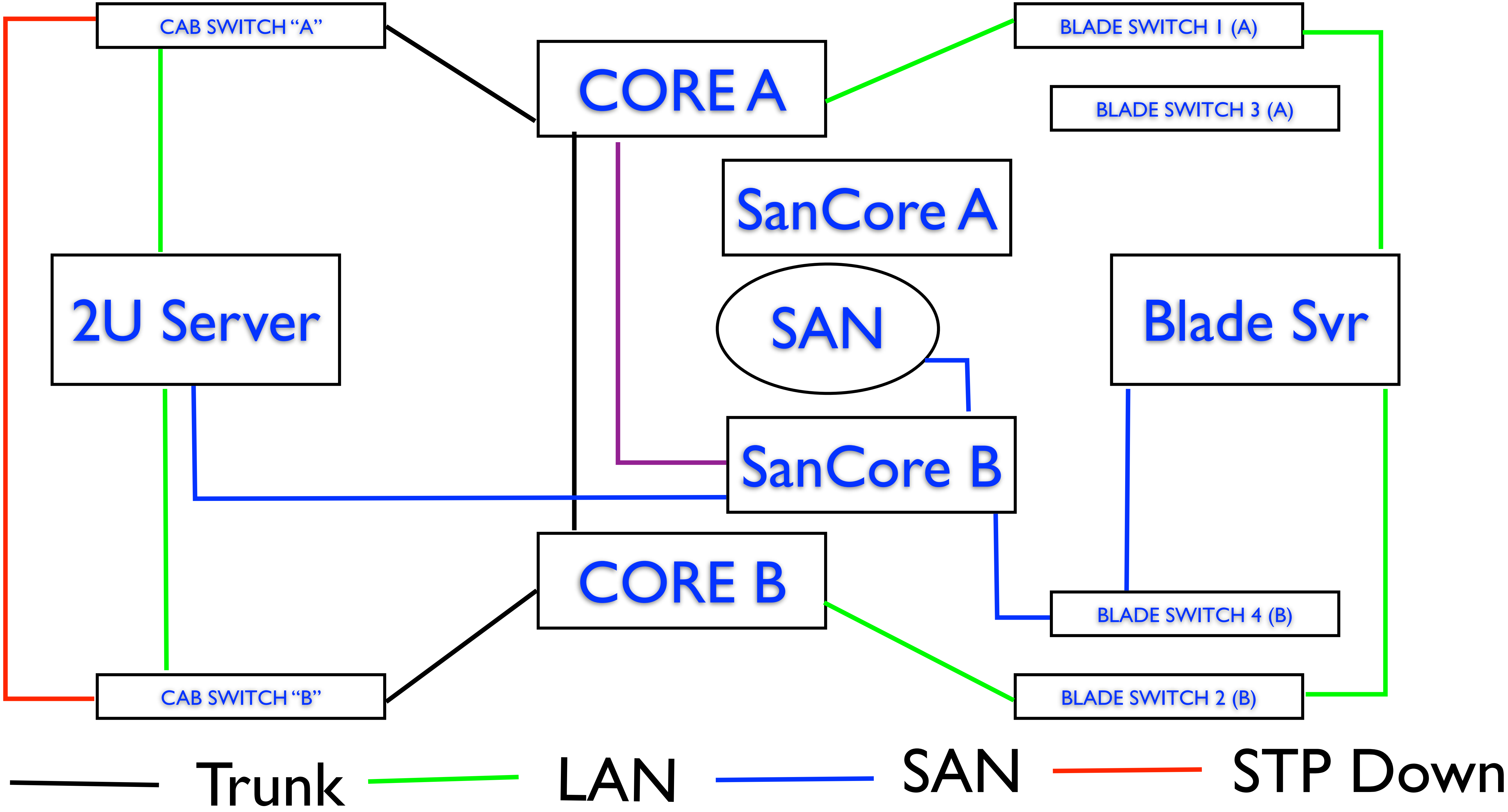
Lather, Rinse, Repeat



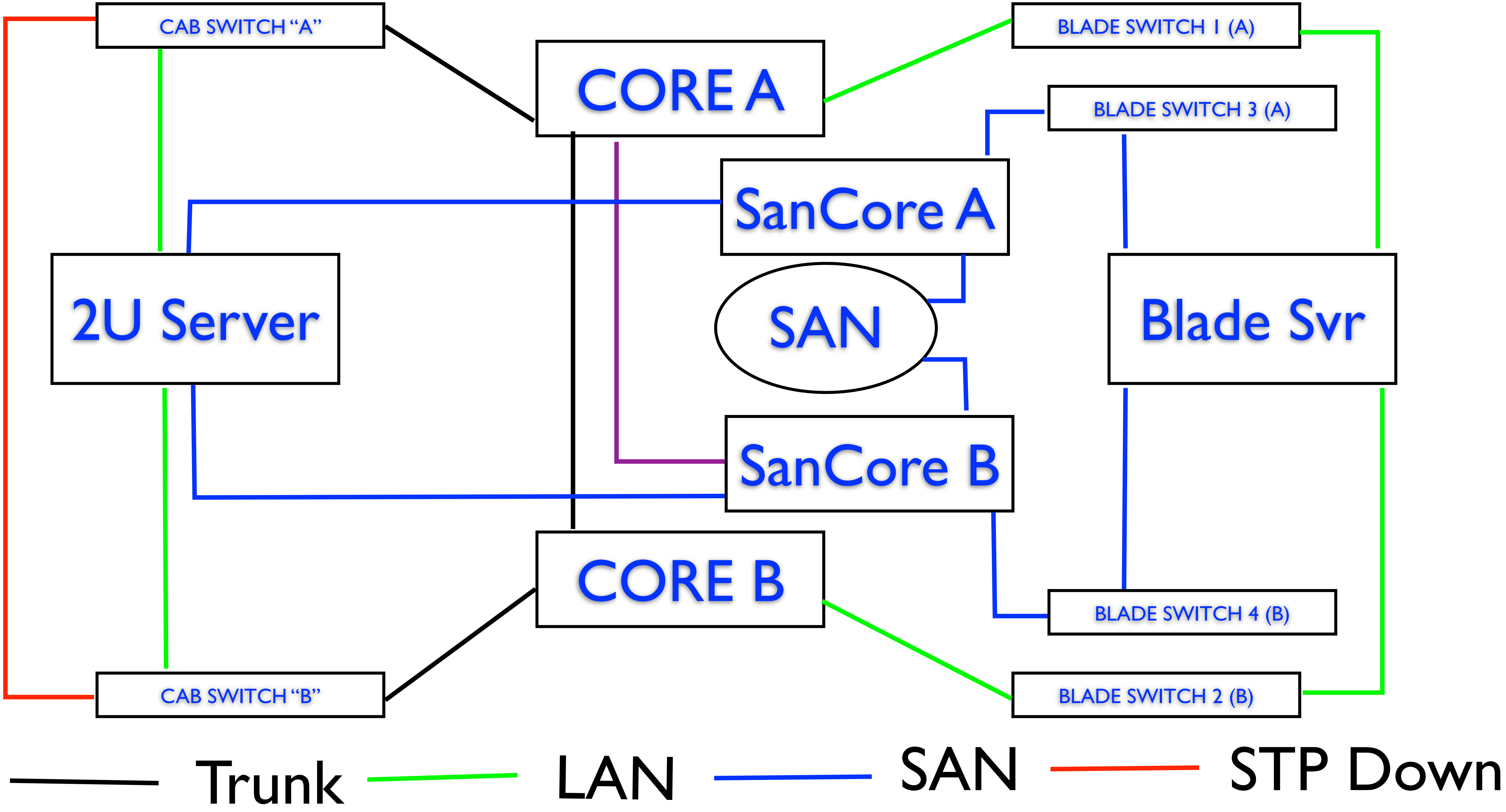
Disable and Disconnect "A" Side SAN Ports



Install New "SanCore A" Switch



Connect All "A" Side Cables To The New SanCore A



CAB SWITCH "A"

CORE A

BLADE SWITCH 1 (A)

BLADE SWITCH 3 (A)

SanCore A

2U Server

SAN

Blade Svr

SanCore B

CORE B

BLADE SWITCH 4 (B)

CAB SWITCH "B"

BLADE SWITCH 2 (B)

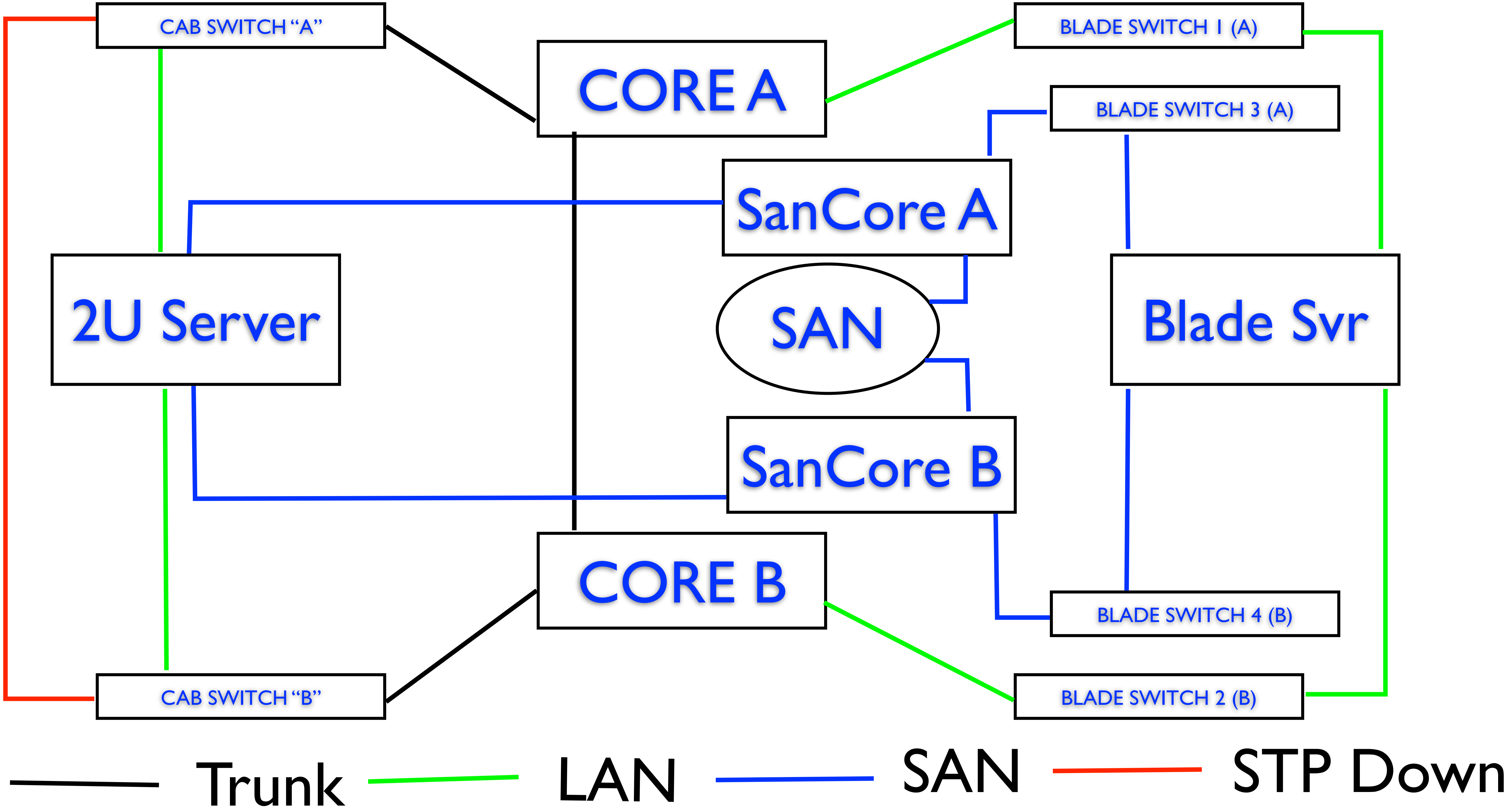
Trunk

LAN

SAN

STP Down

Remove The Temporary Cable



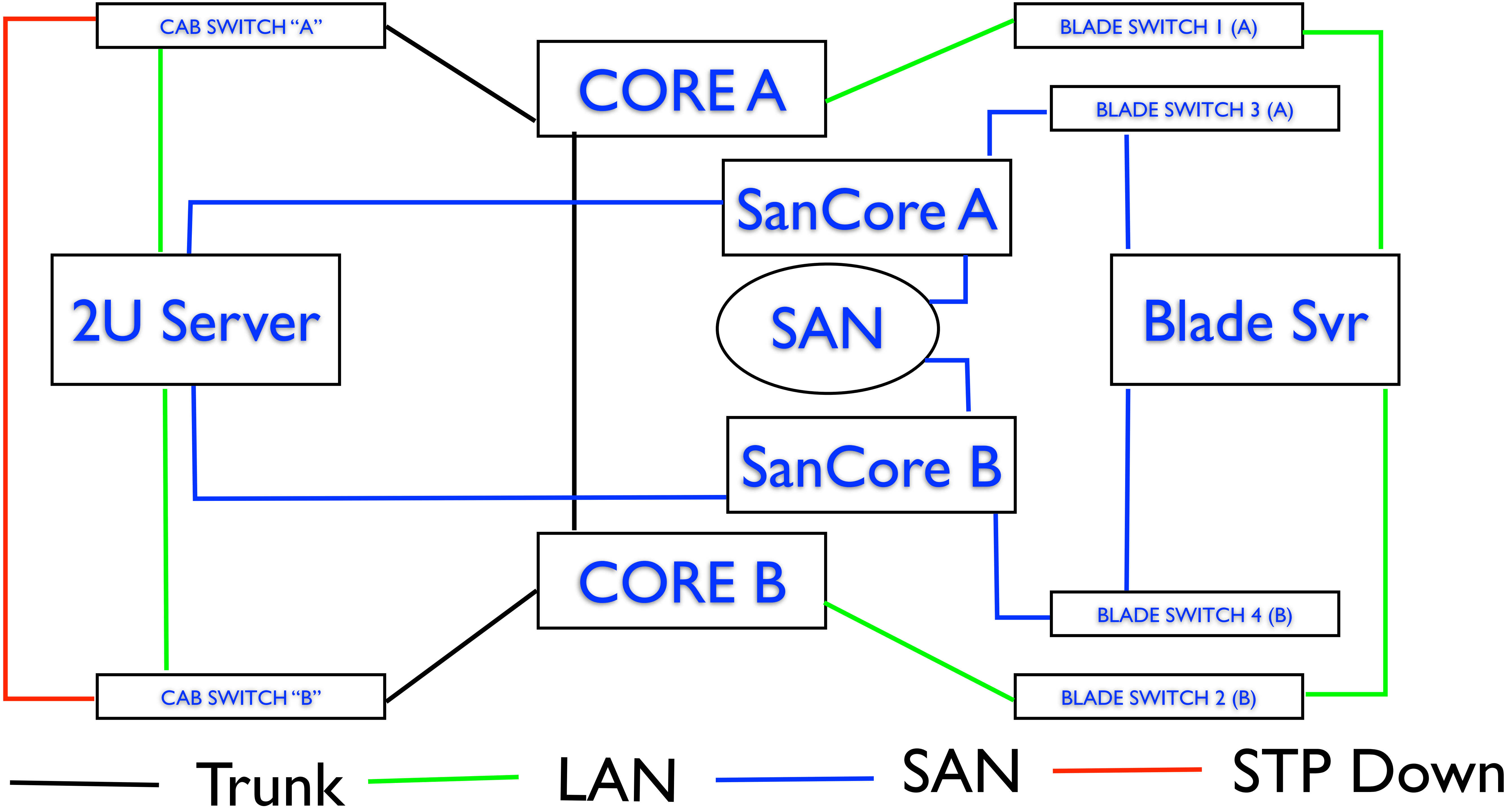
Lather, Rinse, Repeat

- Disable all the "A" side SAN ports on blades, 2Us, and SAN modules
- Everything should seamlessly switch to using the "B" side infrastructure
- Once the "A" side ports have isolated themselves from the Core switch, install the new "A" side "SAN Core", and move all their cables to the new switch

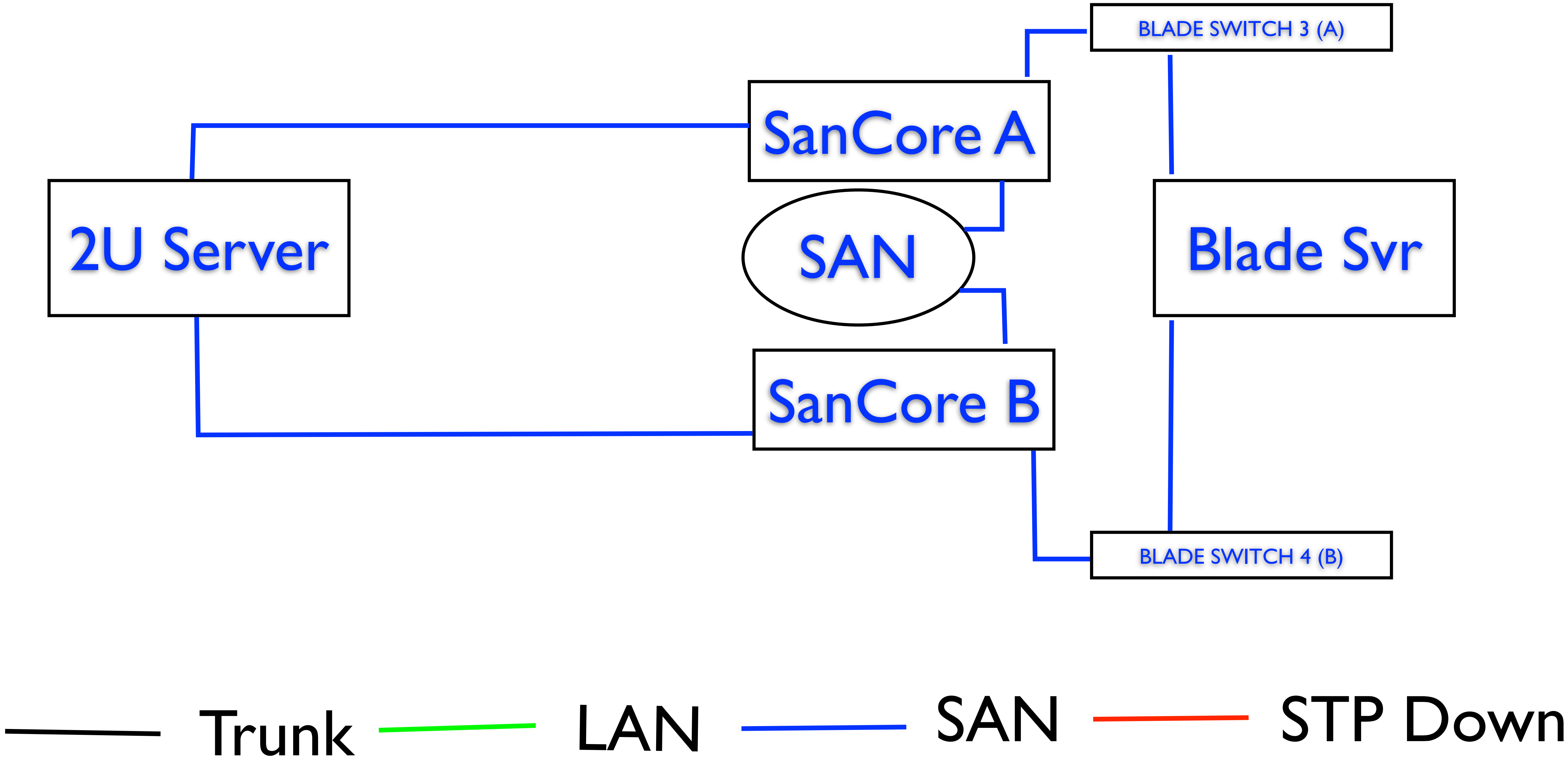
Lather, Rinse, Repeat (part 2)

- You should remove that temporary cable
- (There’s nothing SAN related on the “legacy” network, it’s time to cut the cord)
- Light up the “A” side SAN NICs on all the consumers
- Lo and behold, you just ripped the core networks out of your SAN and (likely) your iSCSI clients didn’t even notice

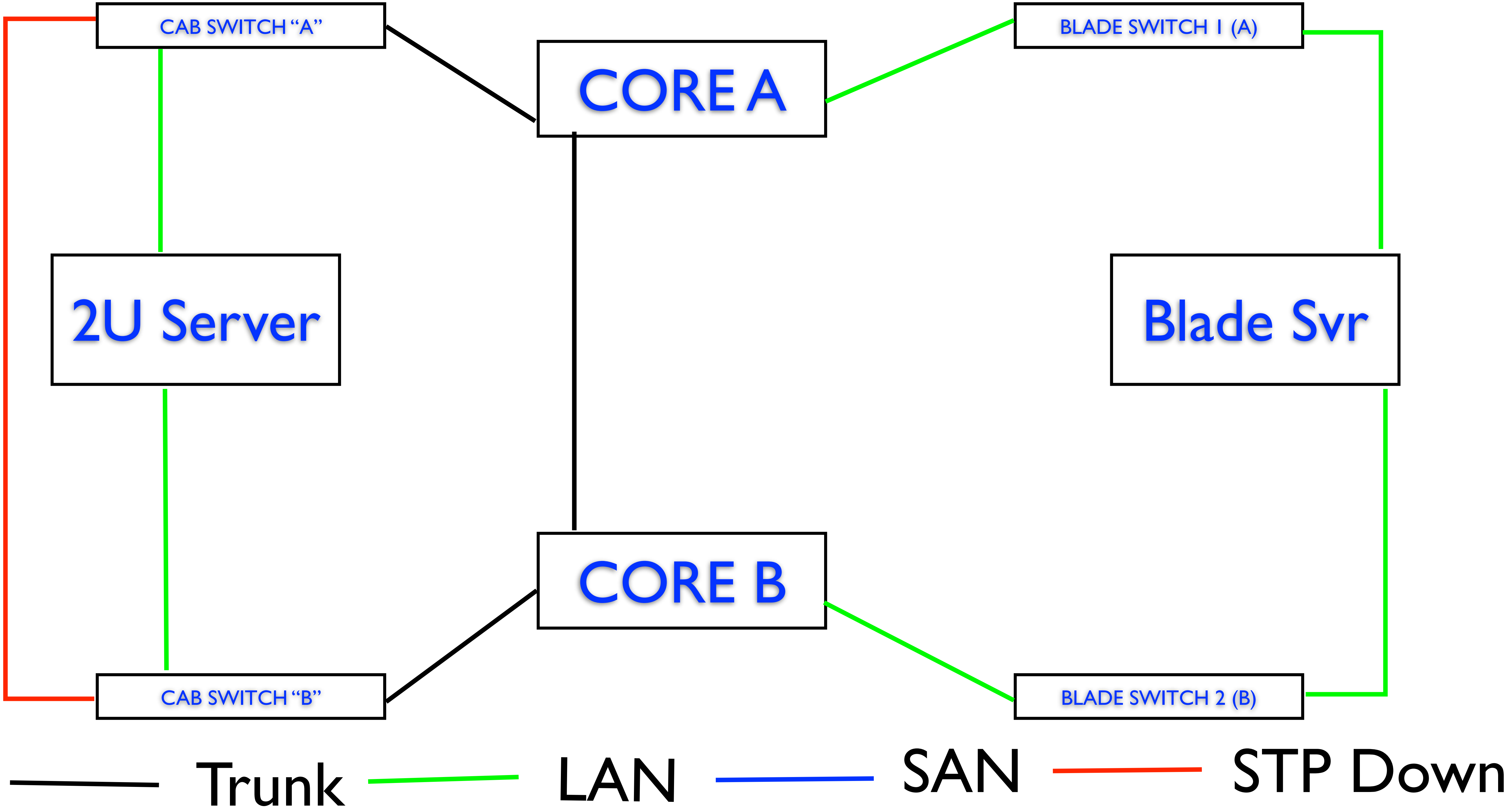
Your Entire Network After The Change



Just The SAN-Related Components



Just The LAN-Related Components



Caution: Results May Prove Addictive

- Once you realize you *can* swap out your core switches without missing a beat, you'll be tempted to do it from time to time
- Done this procedure now four other times since then - replaced the core switches twice, replaced the SAN Core switches twice
- Only dropped the ball once

Dropping The Ball

- How do we NOT “drop the ball”?
- Plan, plan and plan again
- Have some friends of yours read the plan
- Sleep a bit
- Plan some more

The Power of the Whiteboard

- Draw your network diagram on the whiteboard, including every link to every switch (representative samples are fine, obviously)
- For each step in your process, erase/draw lines to represent your changes
- Then, for each step, for every device, ask yourself "what path does this device now use to get from A to B"?
- Be cognizant of "events" you may trigger

Follow That Procedure

- After spending hours working on this procedure, you'll start to have dreams (nightmares) about it
- You'll think you know it inside and out
- You don't.
- When Change Day comes, follow the procedure *exactly as you have written it down already!*
- You will forget some important reason for the order of operations, and you will be very unhappy.

Conclusions

- Again, none of this is rocket-science. It's everything you probably had ever read about redundant networking
- Network administrators, really, have known about how to do this sort of thing forever, but as sysadmins, we don't mess about with it that often ourselves

Conclusions (part 2)

- iSCSI isn't a broadcast laden protocol. Even a largish flat network, used only for iSCSI, probably isn't a big problem for a lot of sites
- Meticulously craft your procedure, and follow it like you might a religious text. If you say to yourself "oh, I can merge steps 17 and 19, and do 18 after", it's likely that you're wrong.
- Find your optimizations of process on the whiteboard, not on the fly.

Questions?

Thanks!

- e-mail: derekb@answers.com or dredd@megacity.org
- slides: http://www.megacity.org/slides/