

# Providing Packet Obituaries

Petros Maniatis

with Katerina Argyraki (Stanford),  
David Cheriton (Stanford),  
Scott Shenker (UCB/ICSI)

Intel **Research**  
Berkeley

# Executive Summary

- This is new work
  - Rant and/or rave, please!
- Problem motivation
  - Integrated delivery status info with every IP packet
  - Inline feedback for actual (not probe) traffic
- Machinery
  - AS-to-AS boundary packet trackers
  - Hop-by-hop per-packet feedback
- Feasibility
  - Decent BW overhead
  - Preliminary HW design

# Introductory Rant

- Everybody wants network-aware application adaptation
  - Control my end-to-end route to get property X
    - NIRA, WRAP, Platypus, etc.
  - Control my FEC to match unfavorable conditions
- ASes specify in SLAs expected guarantees from their
  - Providers
  - Peers
- How does AS A check that those guarantees are met by neighbors B, C, ...?
  - Probing
  - Keep pinging different destinations via each neighbor counting losses and latencies
  - But what if it's not A's immediate neighbor's fault?

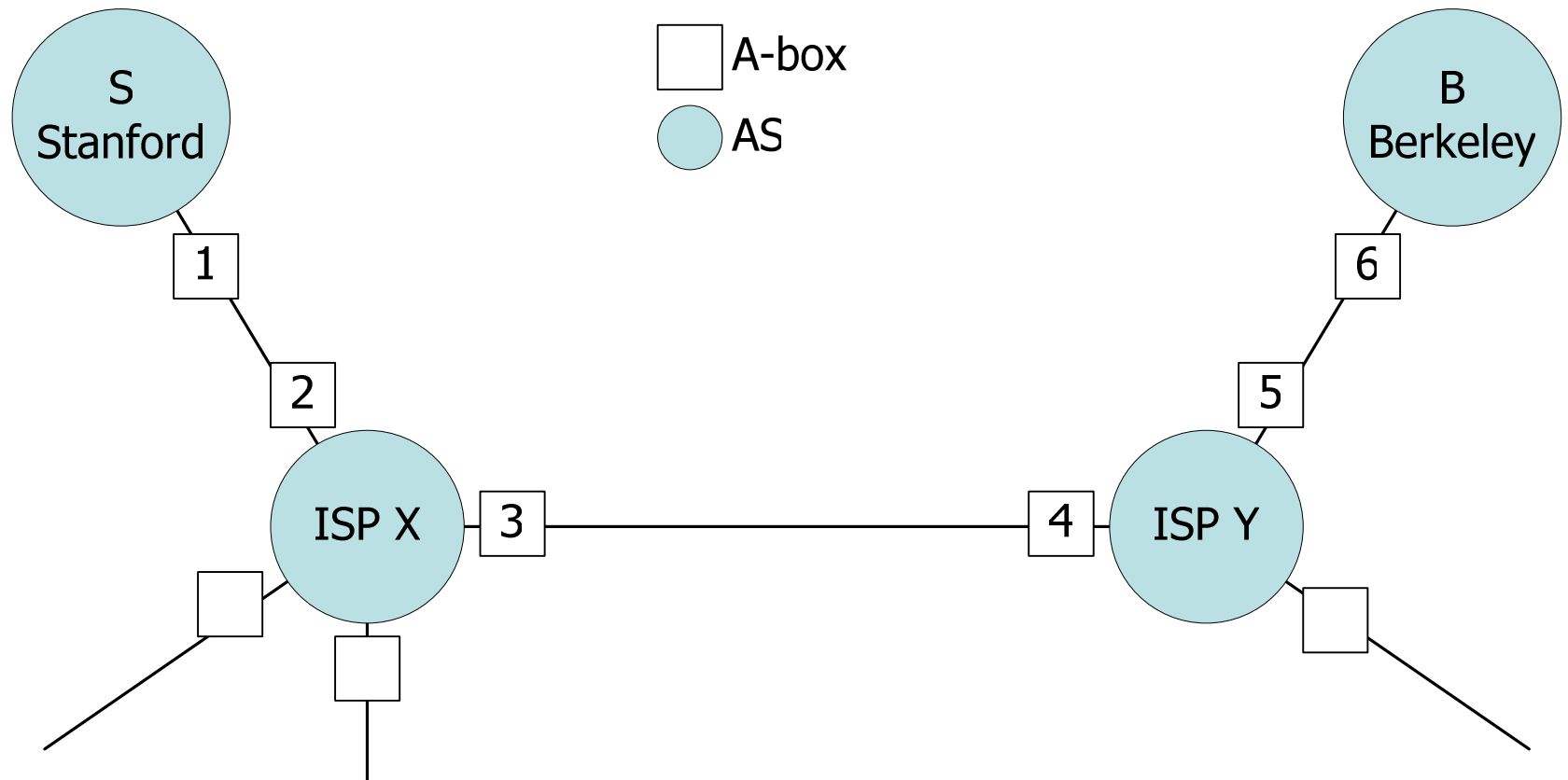
# Probing probing

- Probing is cool
  - I send a packet of some sort to a destination
  - The destination bounces it back to me
  - I count how long it took to go there and back again
- ... but insufficient
  - It measures only how well the probing packets did
  - If the destination does not bounce probe packets, no dice
  - If someone in between filters probe packets or resps., no dice
  - If I care about a packet that isn't a probe packet, no dice
- ... and unwanted by ISPs
  - Increasingly reluctant to let internal topology information go
  - Ping and traceroute are losing ground
- Somebody still cares!
  - entire OSDI '04 paper about collecting, cleaning, and distributing traceroutes

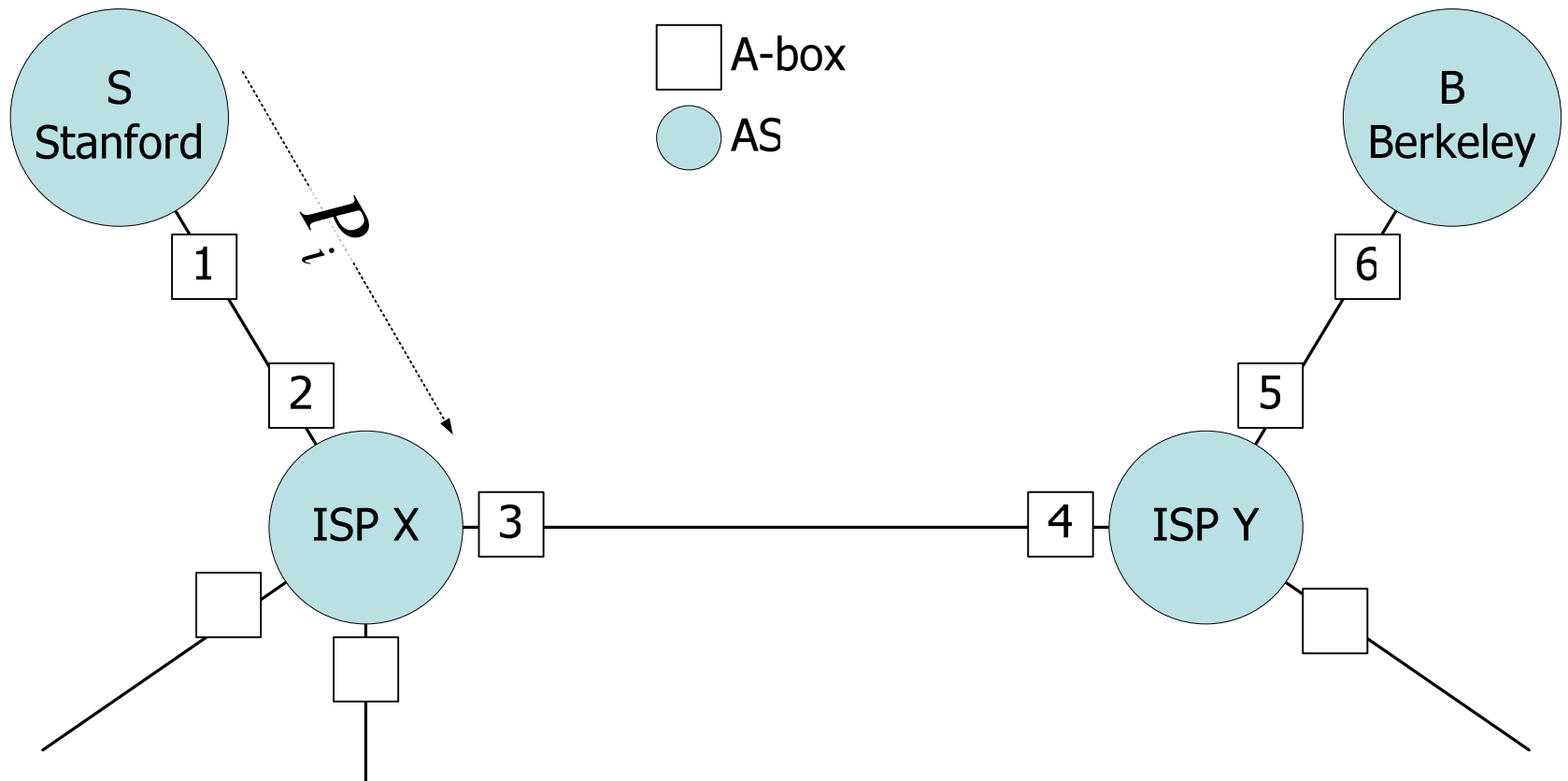
# (Hypo)thesis

- I want to know the fate of every packet I send
  - A packet “accountability” framework
- Solution axes and our choices
  - What: Where a packet dropped (“Obituaries”)
  - To whom: Every hop along packet path
  - Granularity: AS-level (aggregate) information only
  - Who: ASes at their inter-AS boundary links
  - Where: Outside border routers
  - Proactive/Reactive: Proactive
- **ASTRA**
  - An **AS TR**Aceability infrastructure

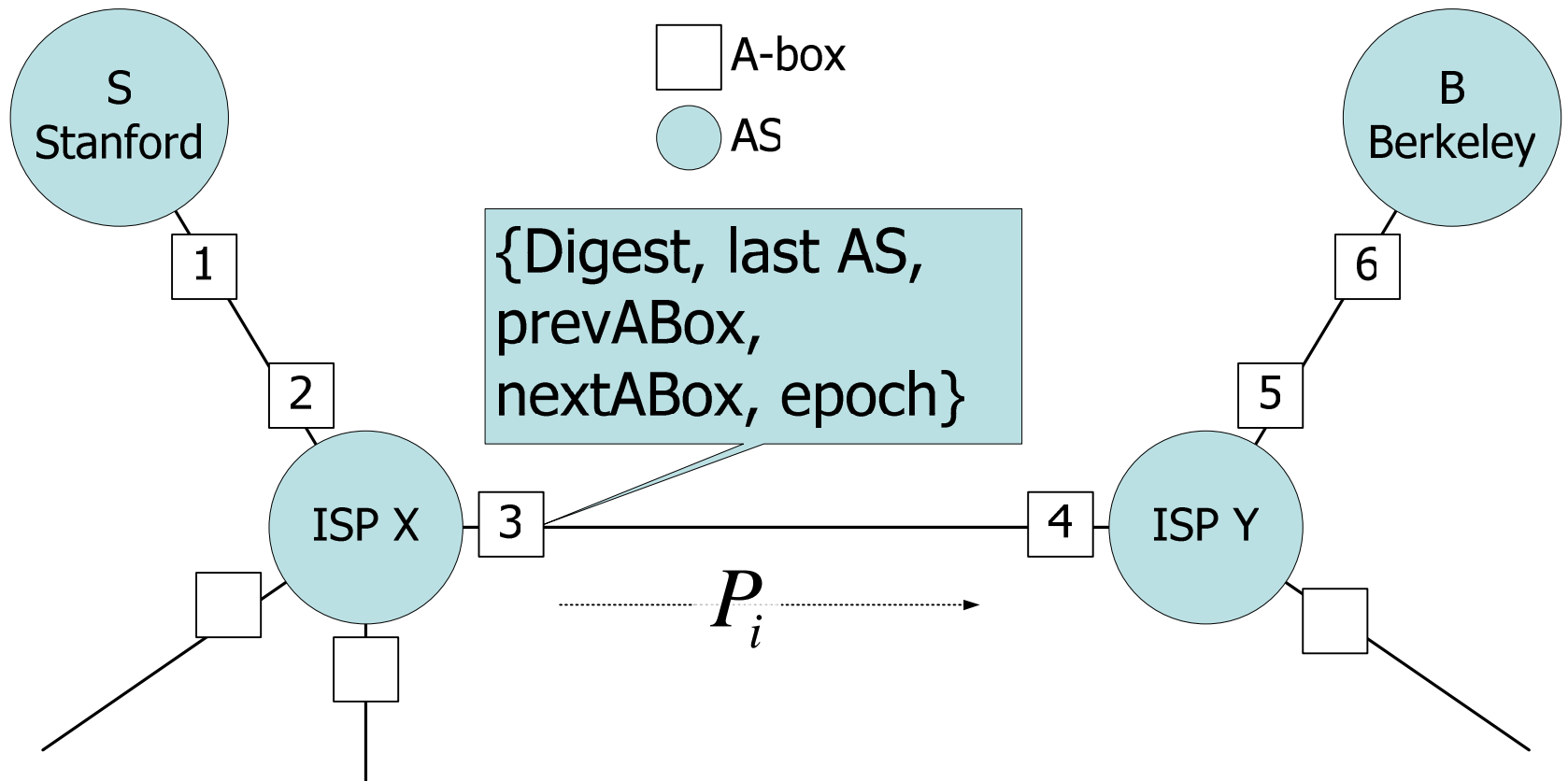
# Basic Framework: Components



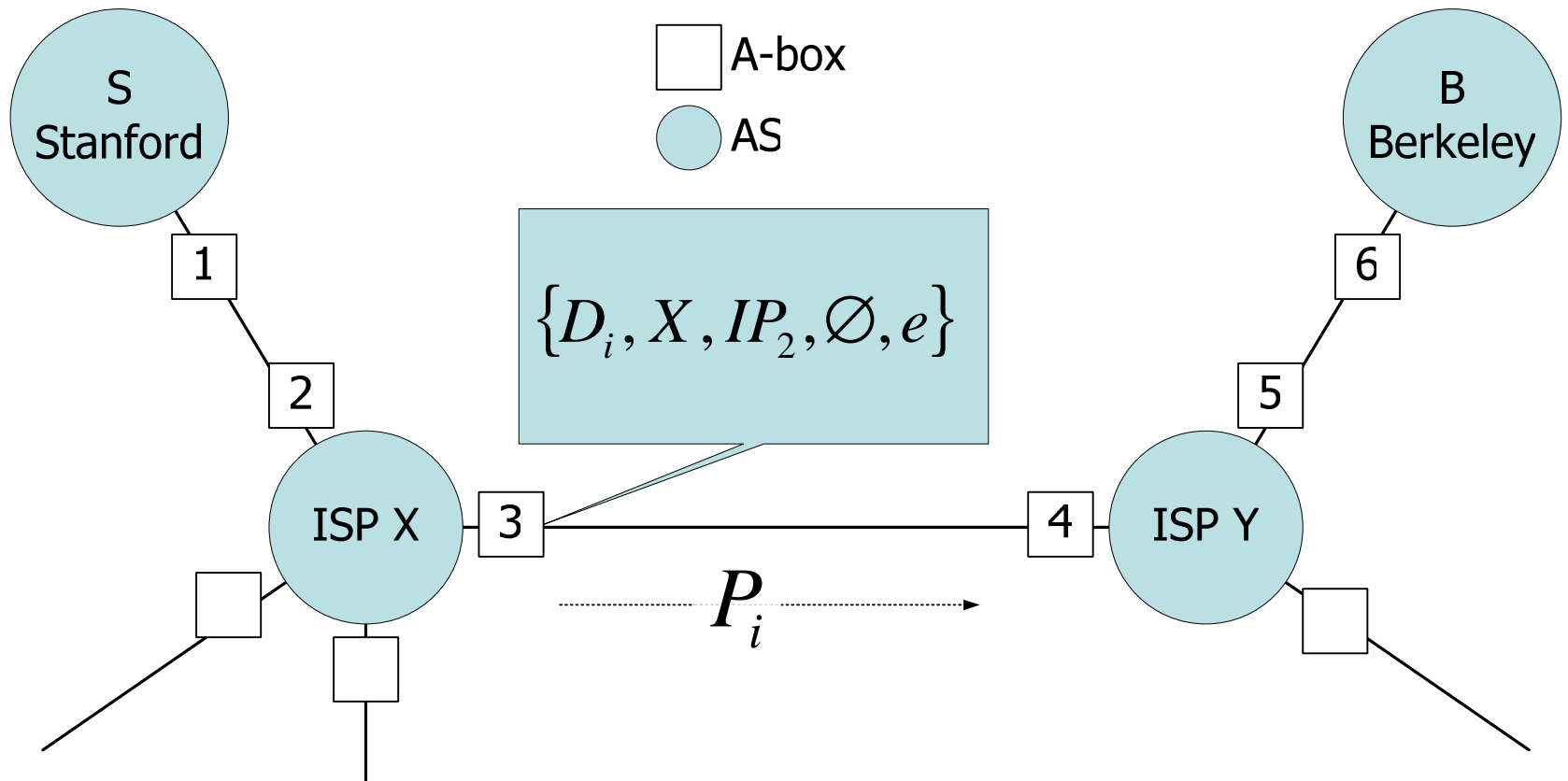
# Basic Framework: Functionality



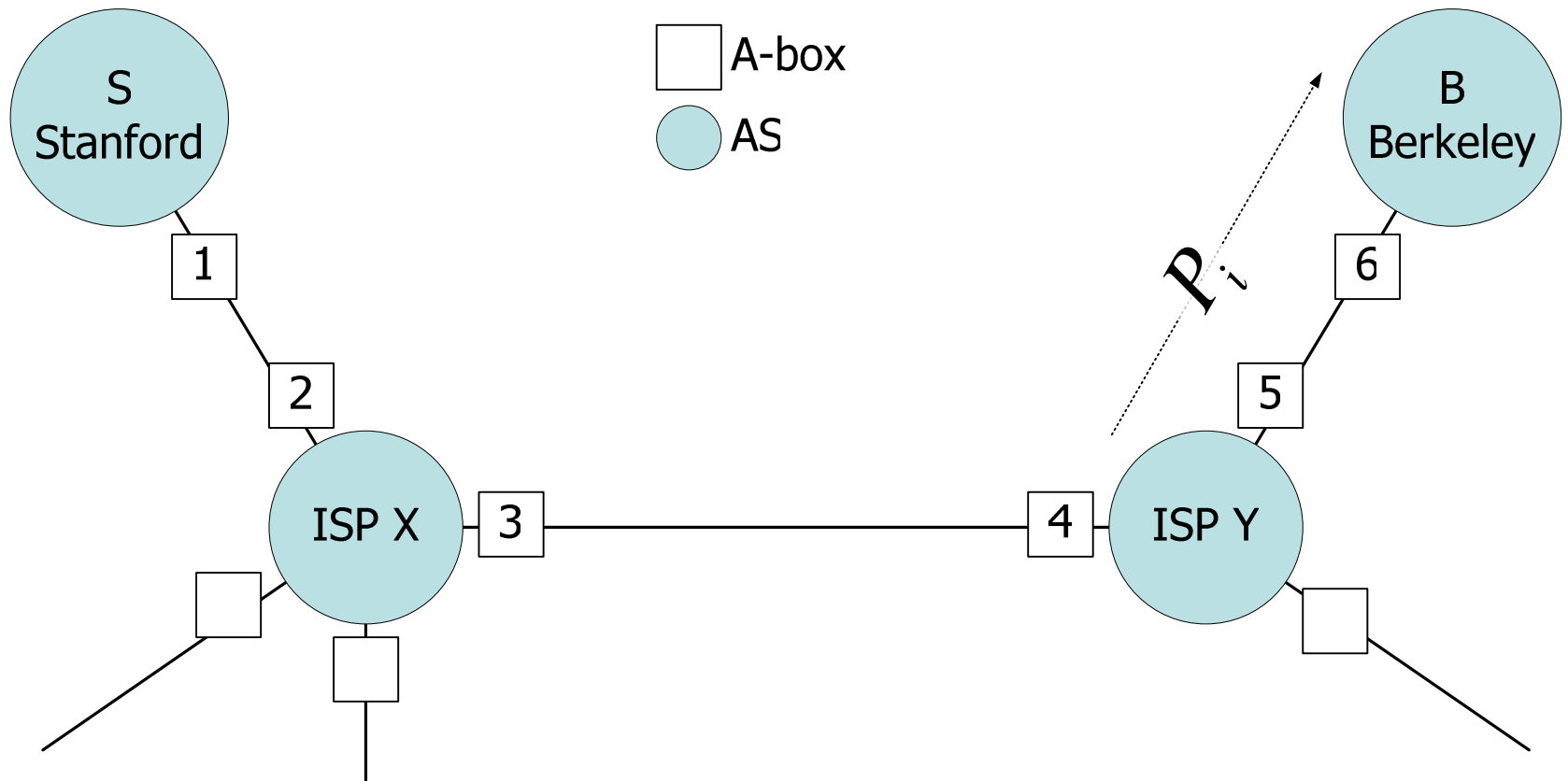
# Basic Framework: Functionality



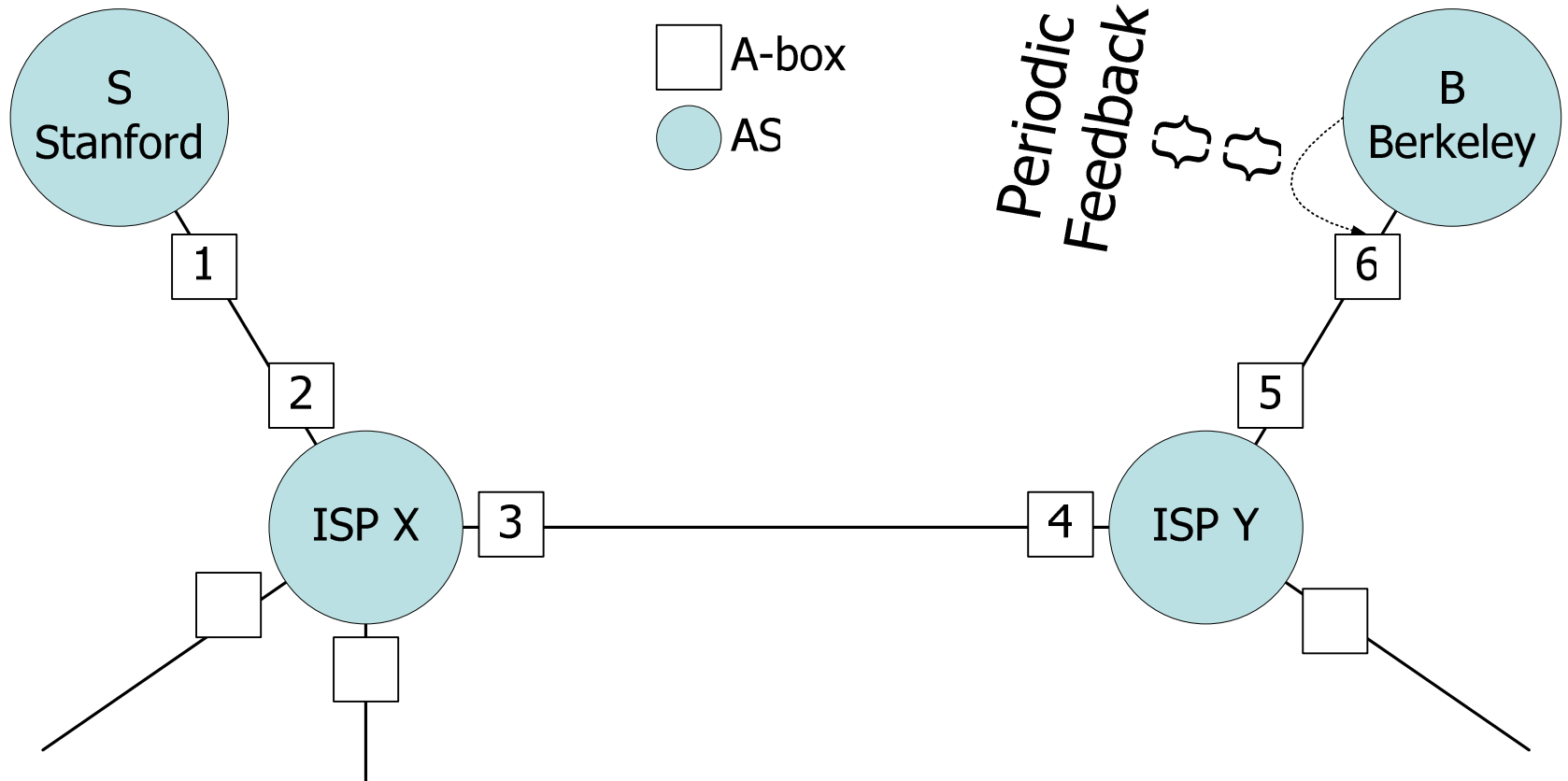
# Basic Framework: Functionality



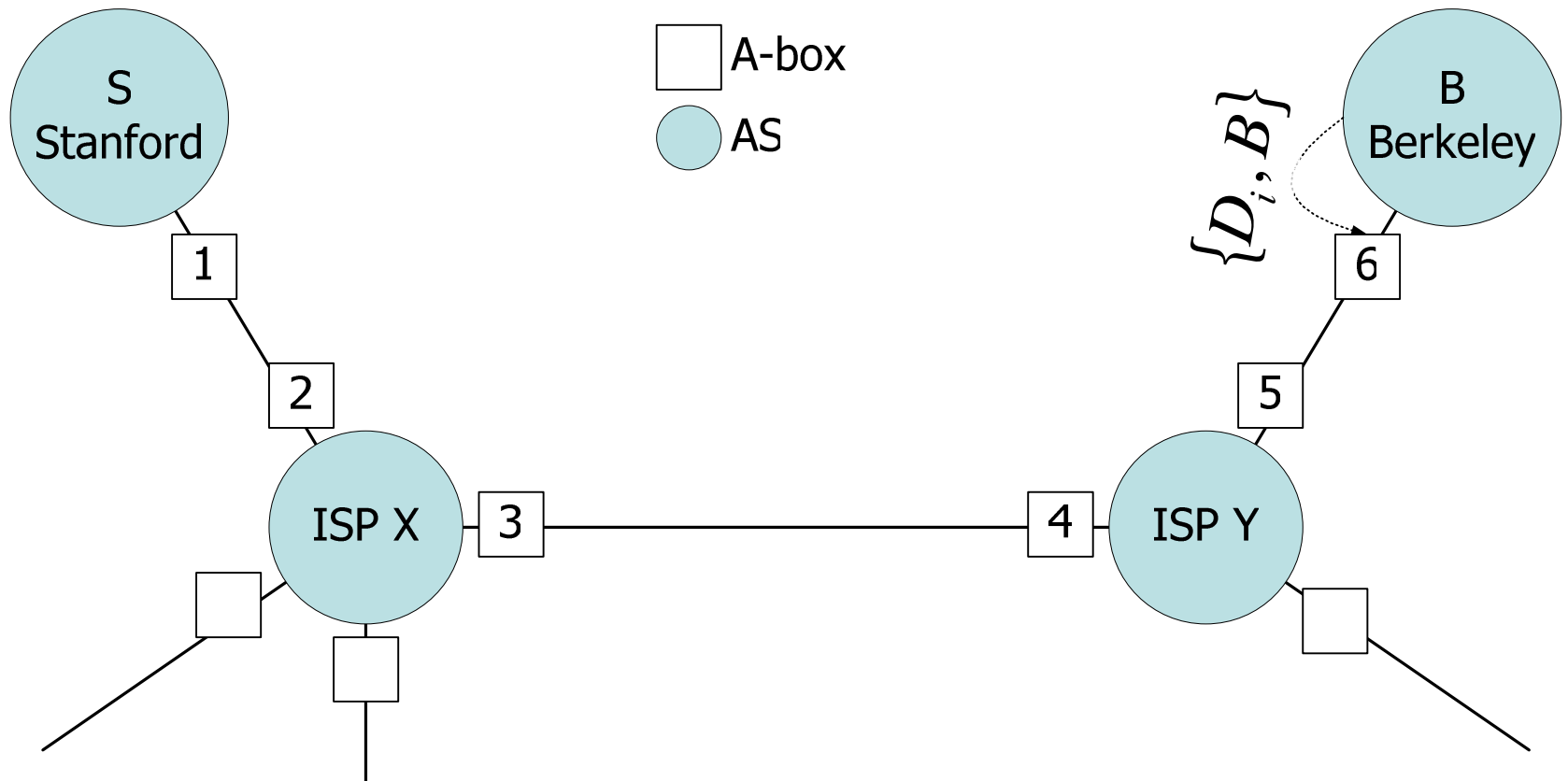
# Basic Framework: Functionality



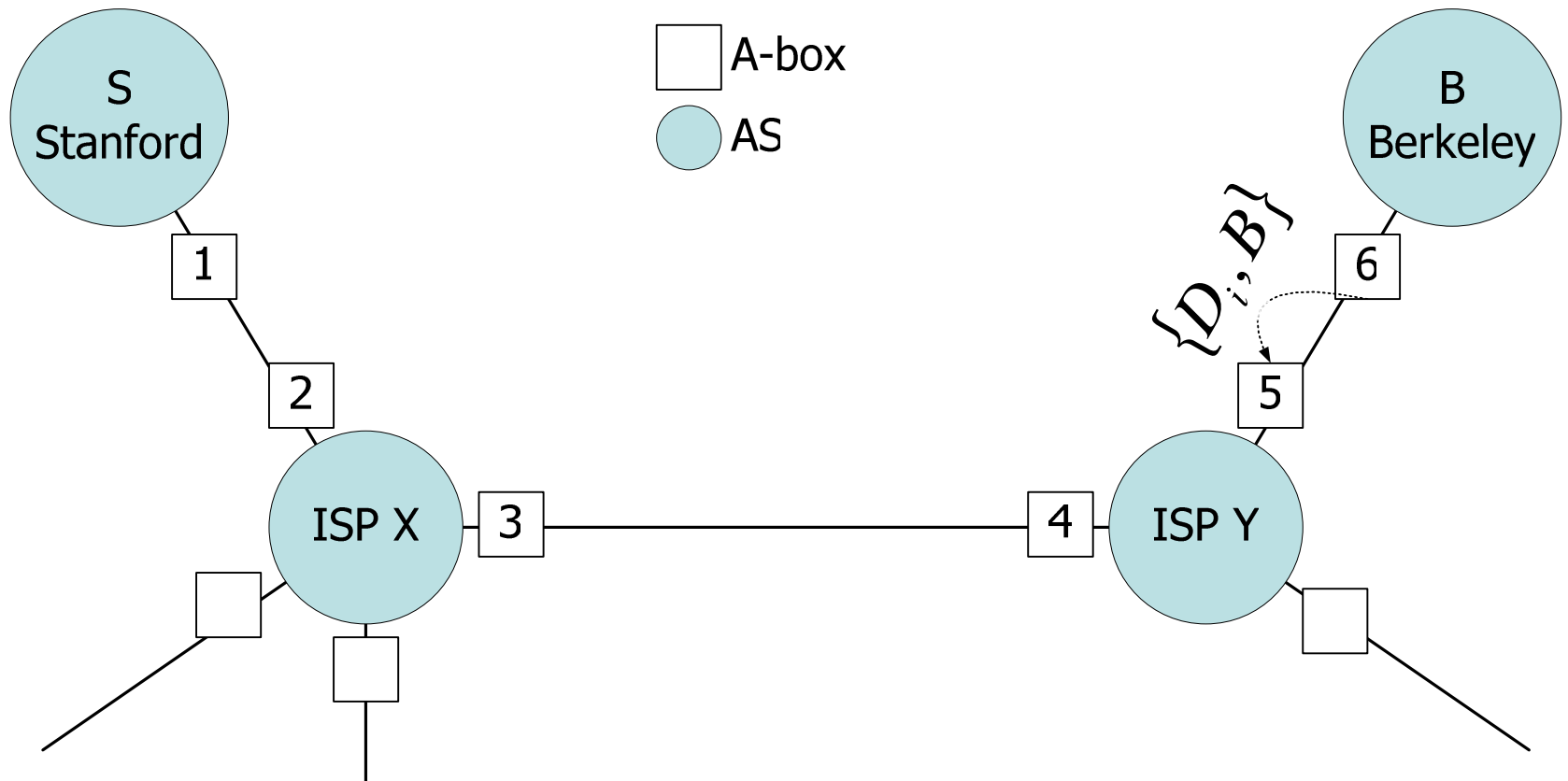
# Basic Framework: Functionality



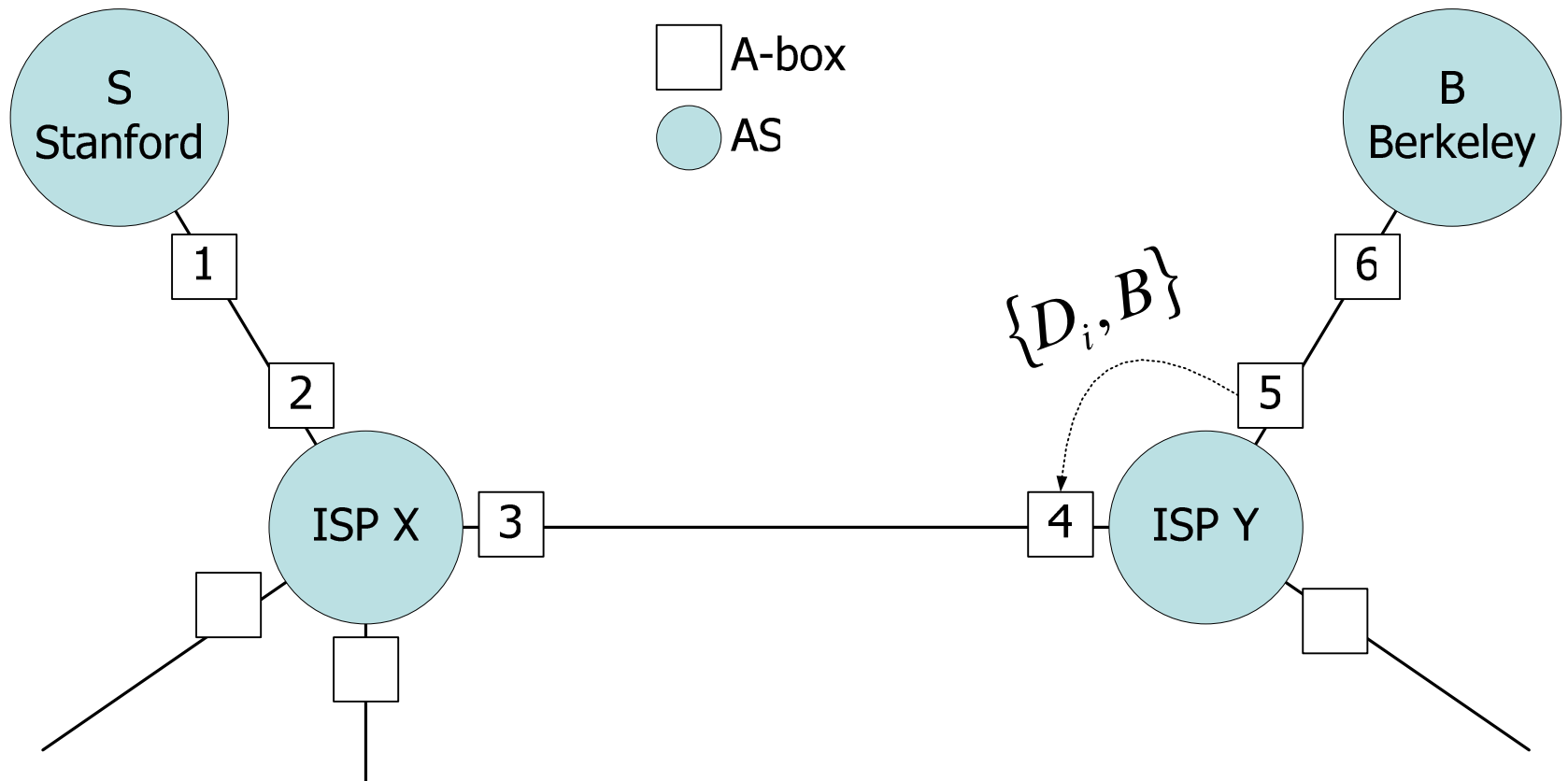
# Basic Framework: Functionality



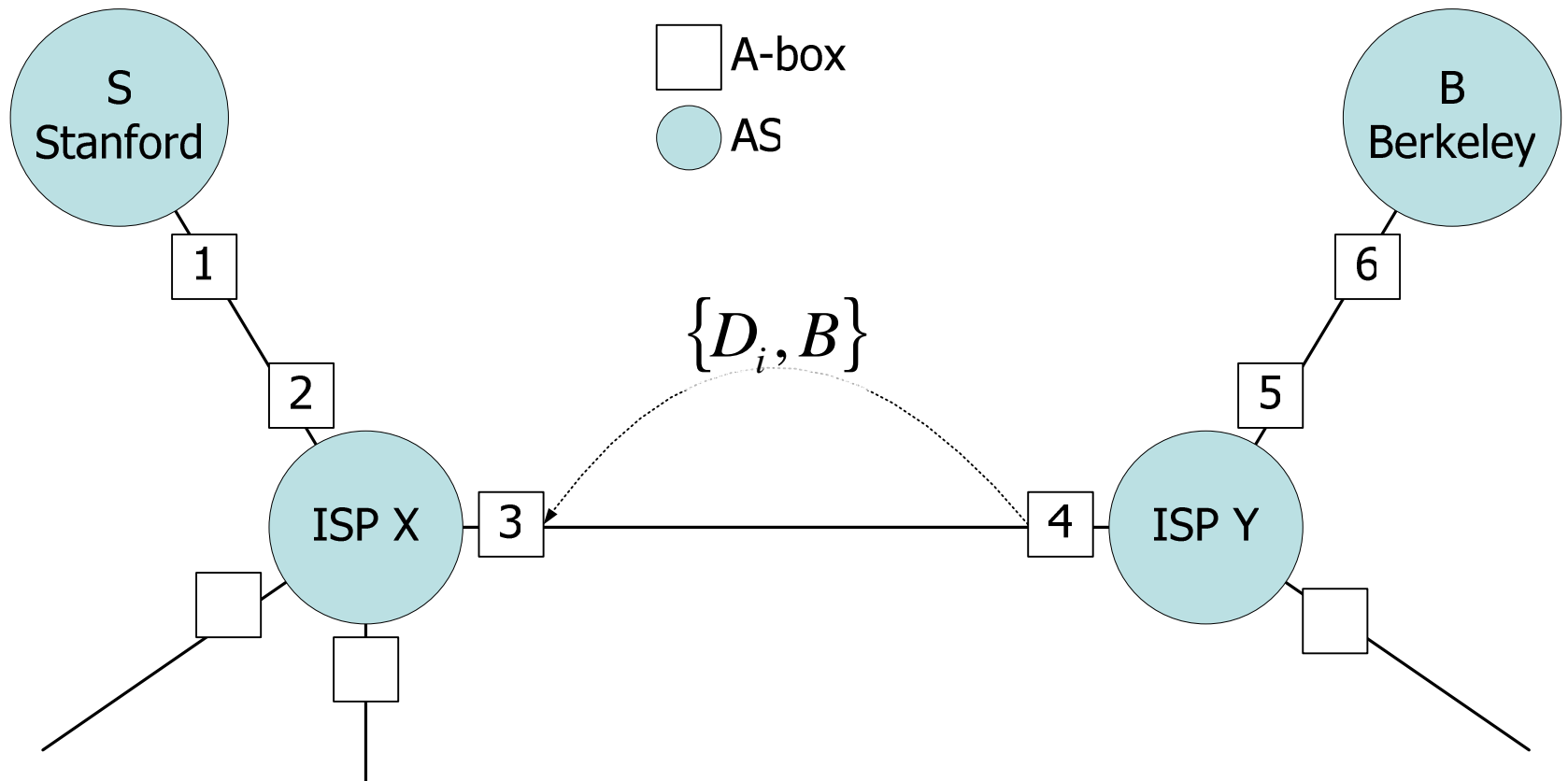
# Basic Framework: Functionality



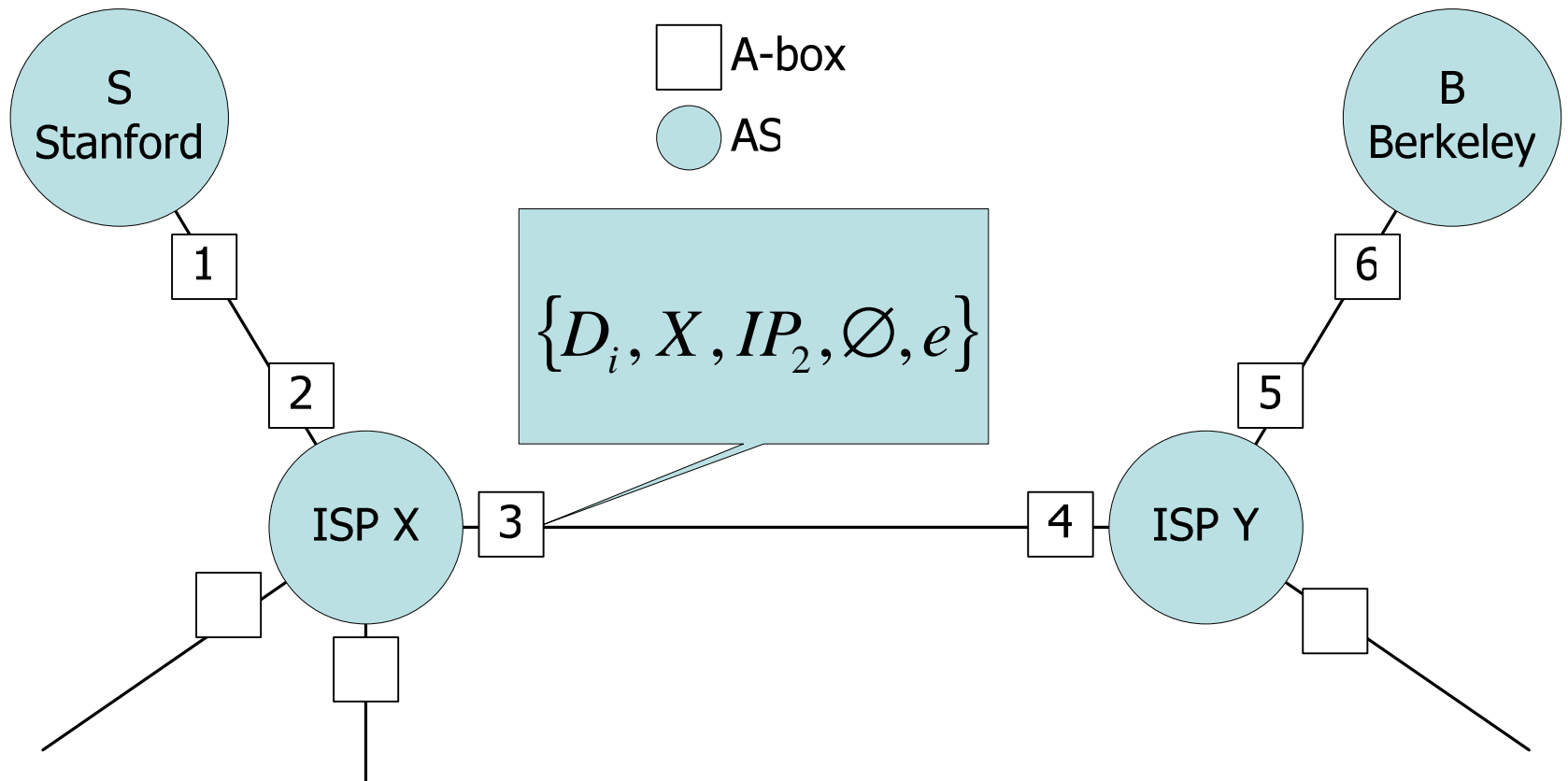
# Basic Framework: Functionality



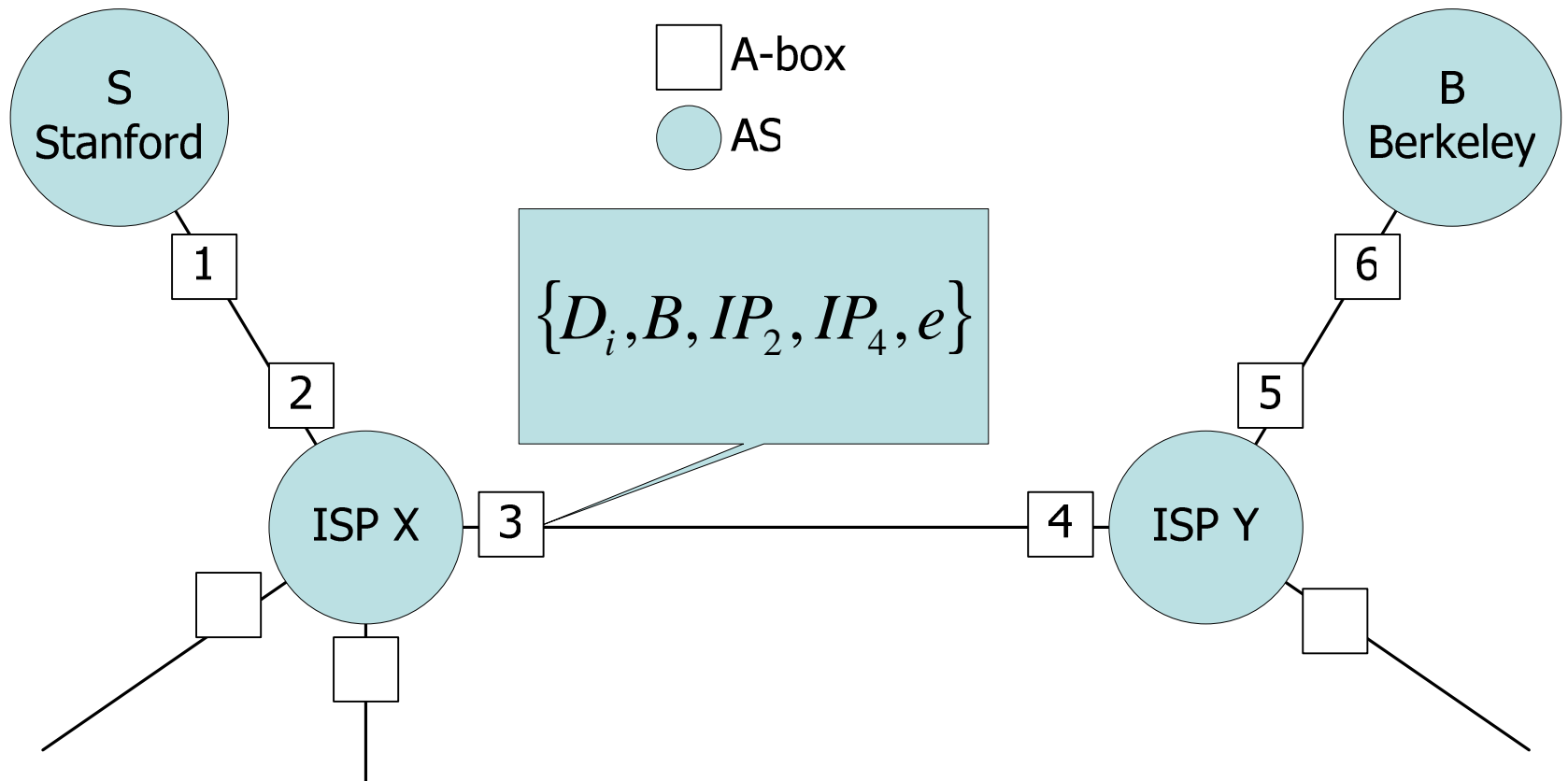
# Basic Framework: Functionality



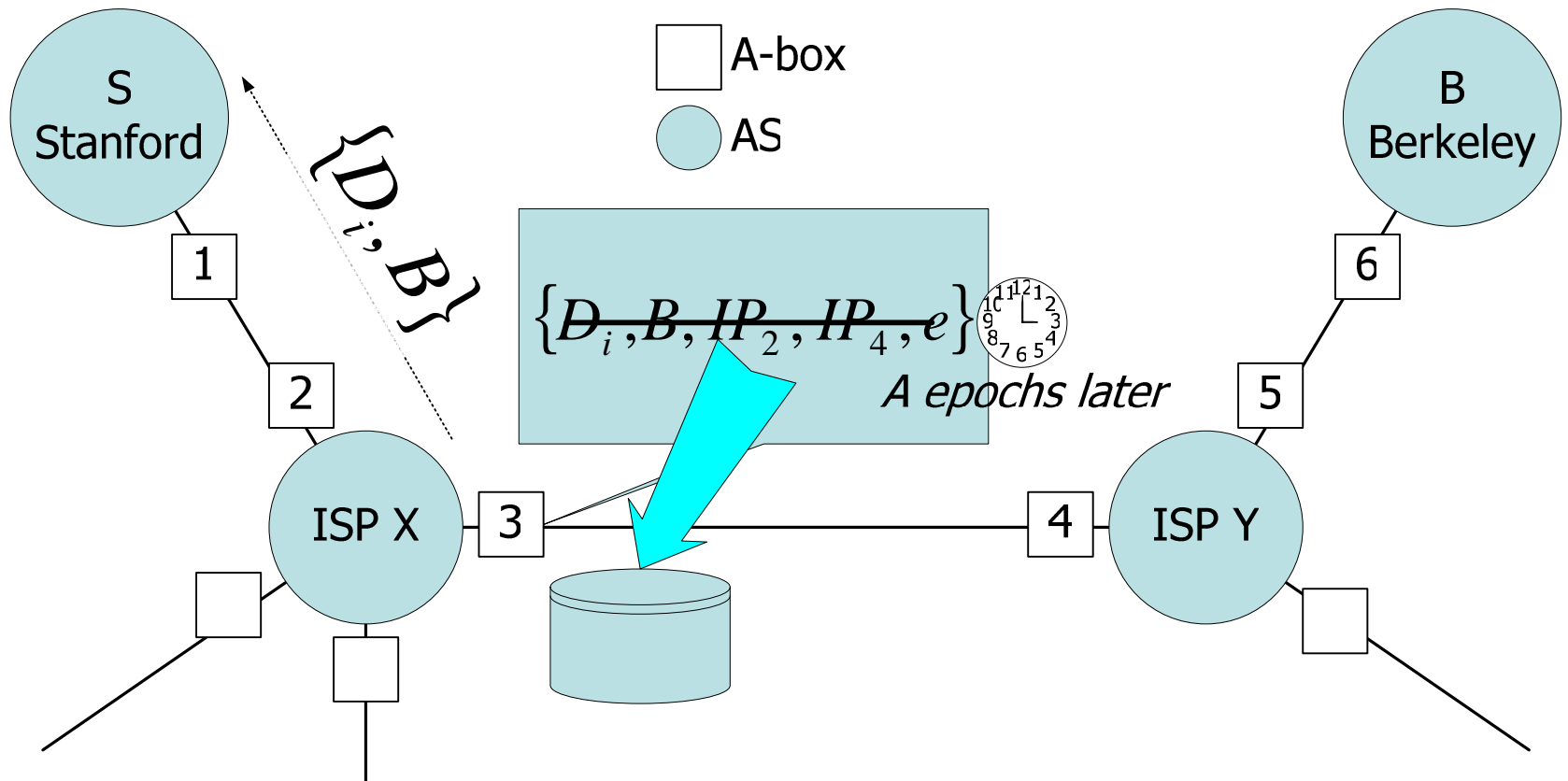
# Basic Framework: Functionality



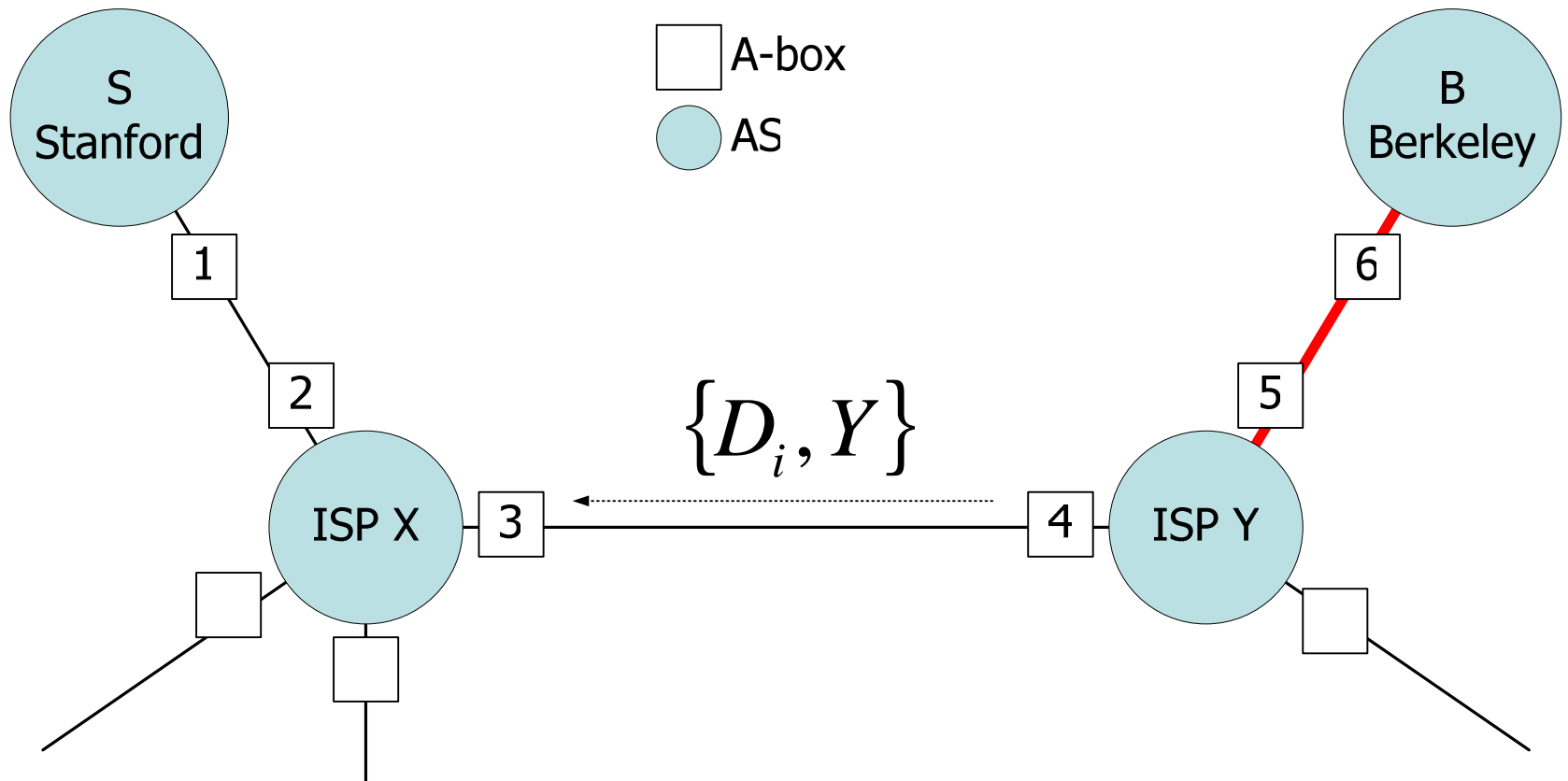
# Basic Framework: Functionality



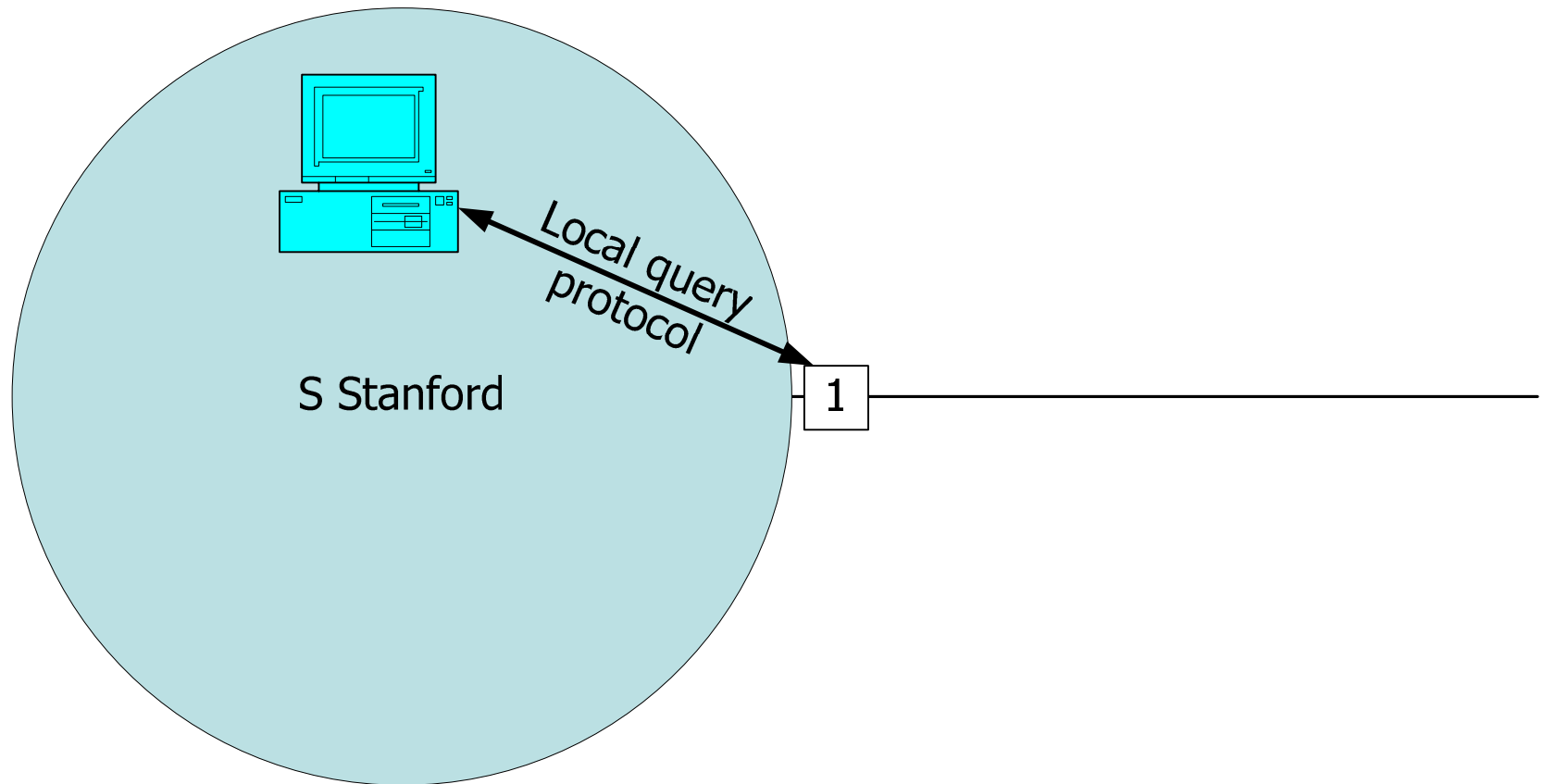
# Basic Framework: Functionality



# Basic Framework: Functionality

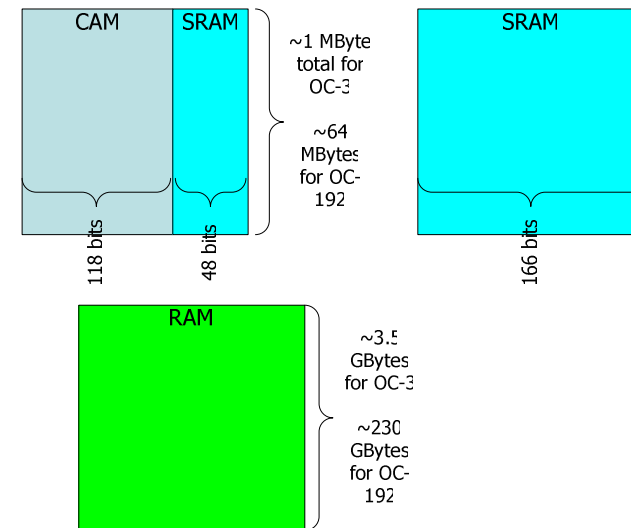


# Basic Framework: Intra AS Interface



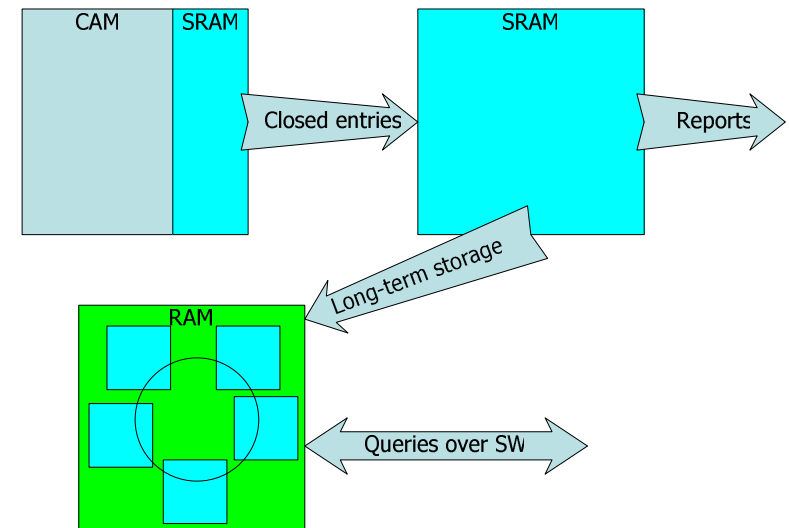
# Basic Framework: Machinery

- CAM and SRAM for short term state
  - CAM (content addressable memory) for indexable fields
    - Digest, previous A-box, epoch counter
  - SRAM for payload
    - Last AS, next A-box, etc.
- SRAM as intermediate buffer
  - For reporting
- RAM for “long” term state



# Basic Framework: Machinery

- New entries created in CAM+SRAM
- CAM+SRAM updated when feedback received
- Report created in SRAM buffer (per destination) and sent out directly
- Entire SRAM buffer transferred to RAM circular buffer
- Long-term state queried via software

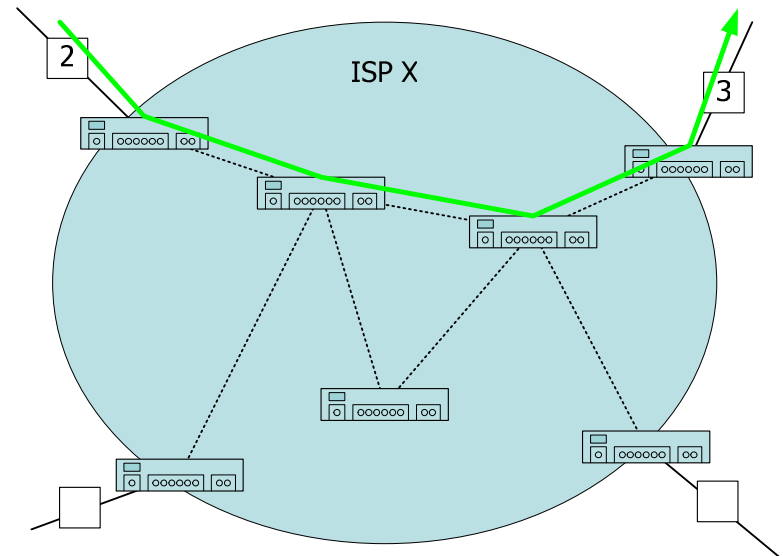


# A bit of detail

- Feedback routing
  - How do I route backwards?
- Discovery
  - What if not everyone runs ASTRA?

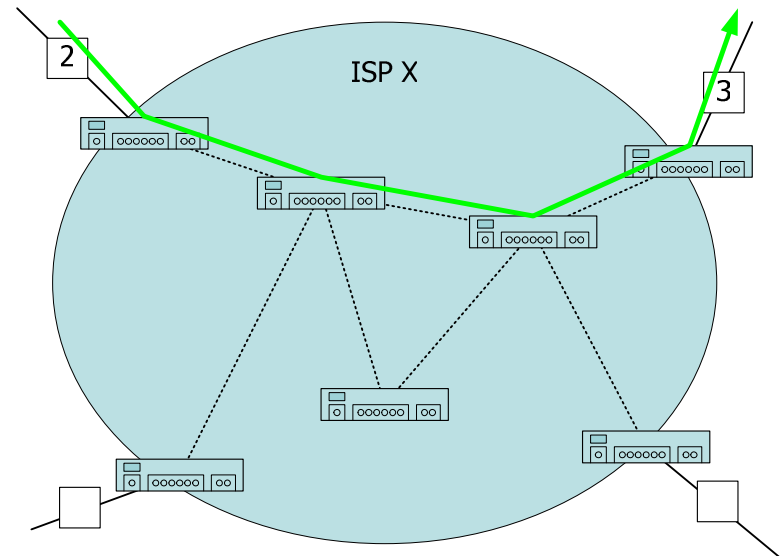
# Feedback Routing

- Within an AS, which A-box should I send my feedback to?
  - Ingress point disambiguation
  - Tackled before, off-line



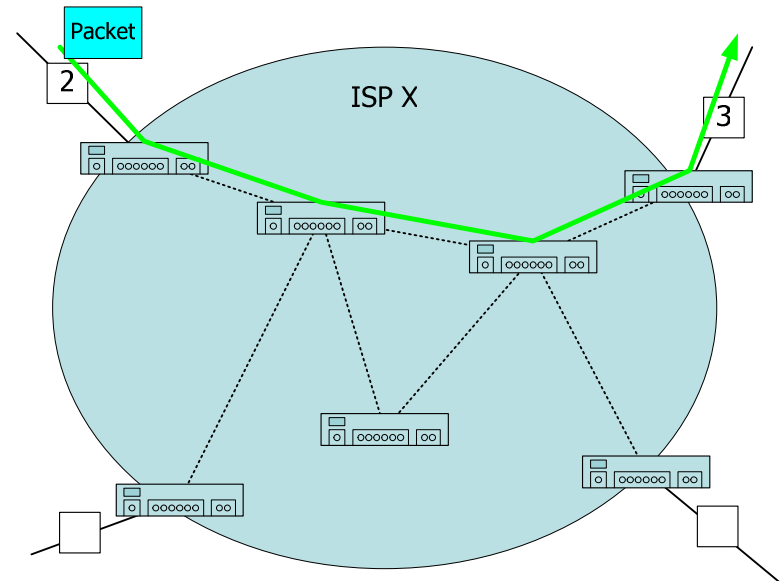
# Feedback Routing

- Within an AS, which A-box should I send my feedback to?
  - Ingress point disambiguation
  - Tackled before, off-line
- Our (brute) approach
  - Packet encapsulation within AS



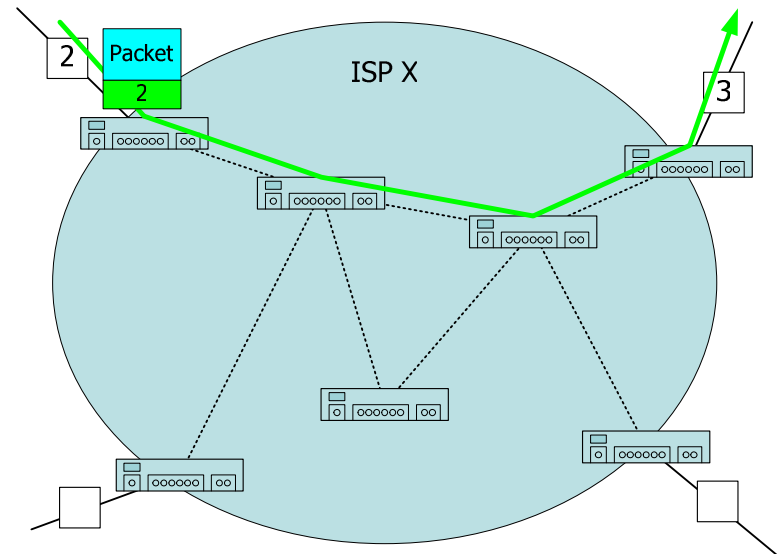
# Feedback Routing

- Within an AS, which A-box should I send my feedback to?
  - Ingress point disambiguation
  - Tackled before, off-line
- Our (brute) approach
  - Packet encapsulation within AS



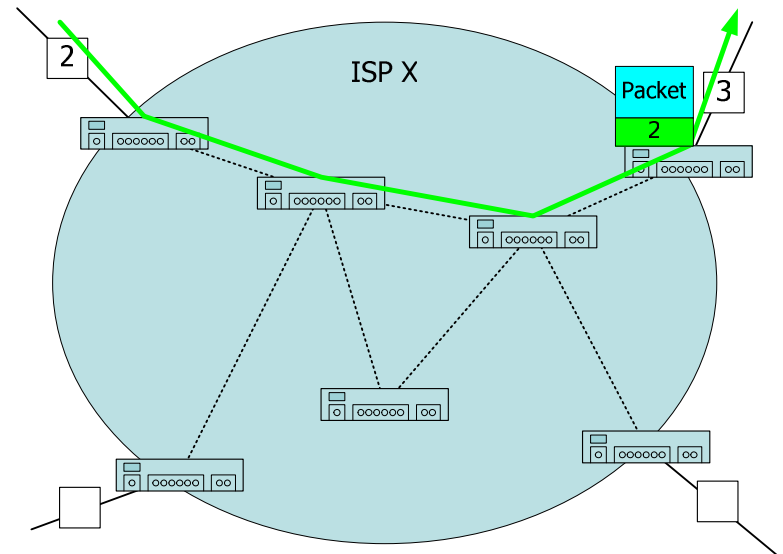
# Feedback Routing

- Within an AS, which A-box should I send my feedback to?
  - Ingress point disambiguation
  - Tackled before, off-line
- Our (brute) approach
  - Packet encapsulation within AS



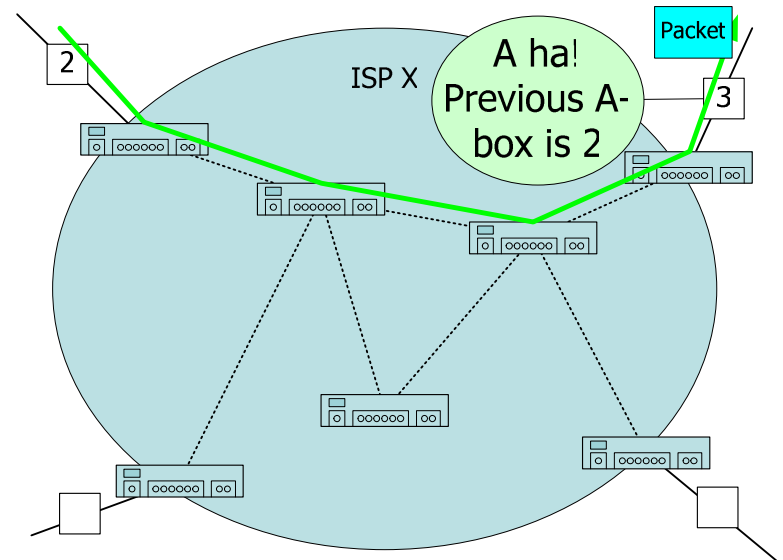
# Feedback Routing

- Within an AS, which A-box should I send my feedback to?
  - Ingress point disambiguation
  - Tackled before, off-line
- Our (brute) approach
  - Packet encapsulation within AS



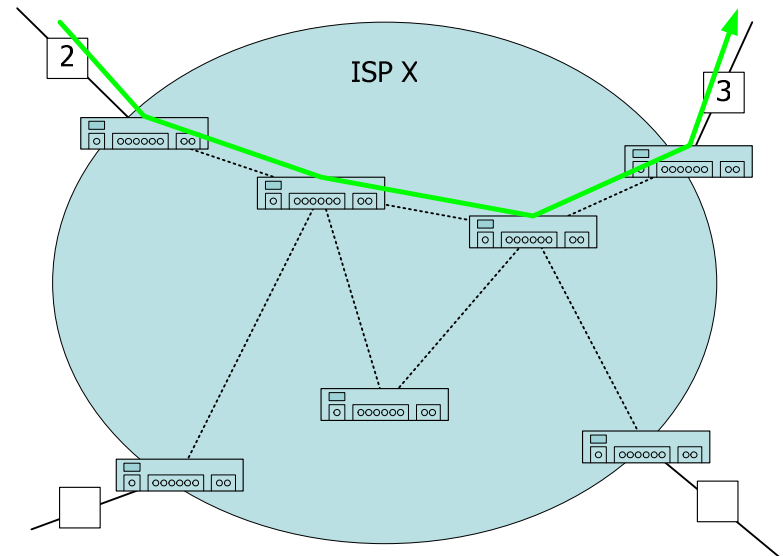
# Feedback Routing

- Within an AS, which A-box should I send my feedback to?
  - Ingress point disambiguation
  - Tackled before, off-line
- Our (brute) approach
  - Packet encapsulation within AS



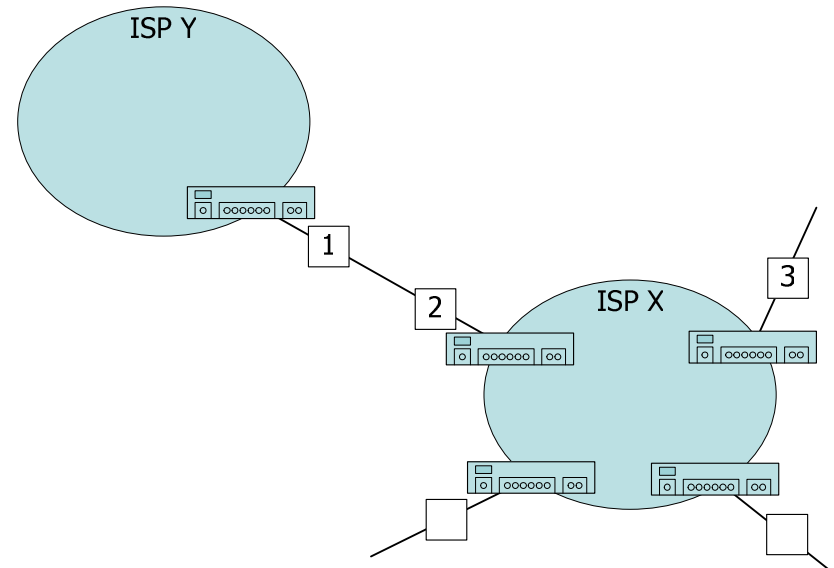
# Feedback Routing

- Within an AS, which A-box should I send my feedback to?
  - Ingress point disambiguation
  - Tackled before, off-line
- Our (brute) approach
  - Packet encapsulation within AS
- Other approaches exist
  - Companion packet
  - Reverse forwarding tables
  - Others?



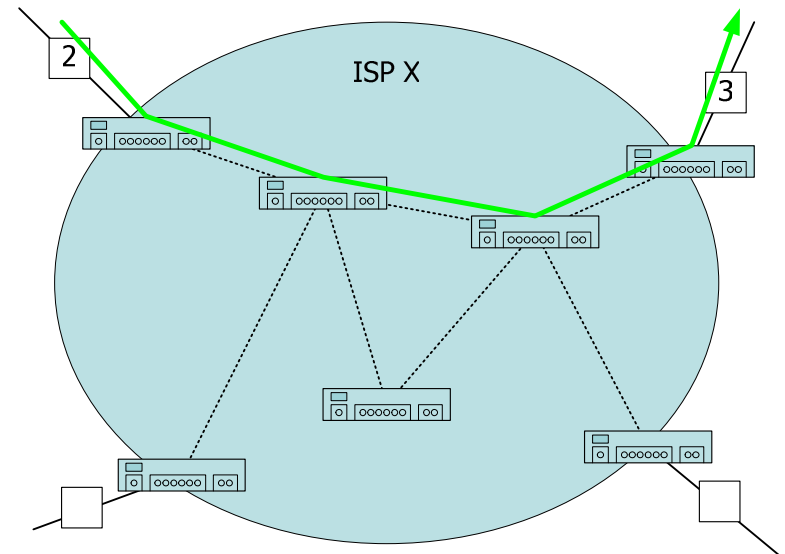
# Discovery

- Last slide was about intra-AS
- Inter-AS question is:
  - What's the address of the A-box across boundaries that I should forward feedback to?
- When the AS on the other side of the link runs ASTRA
  - Easy job: part of configuration
- What if the AS on the other side of the link does not run ASTRA?



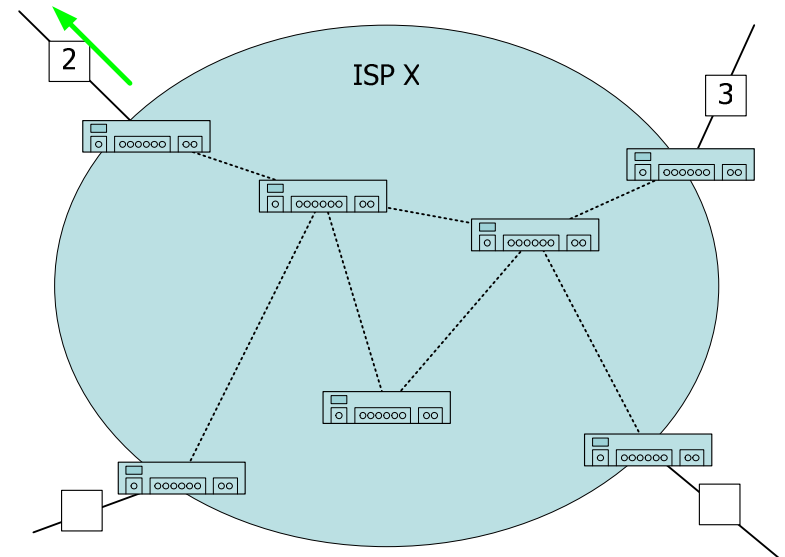
# Discovery with Partial Deployment

- I want to announce myself to every other AS
- Idea: split all ASes among all local A-Boxes, based on who is on the egress towards that AS
  - Pick a prefix per AS
  - See where I'd route it
  - If I'd route it through my AS, don't announce
  - If I'd route it outwards, send my AS number and my IP address
- If I receive an announcement
  - Respond with my own AS number and IP address
- Keep responses from ASes that are the "closest" based on the BGP tables
- Low traffic requirements
  - About 9Mbps over the **entire Internet**



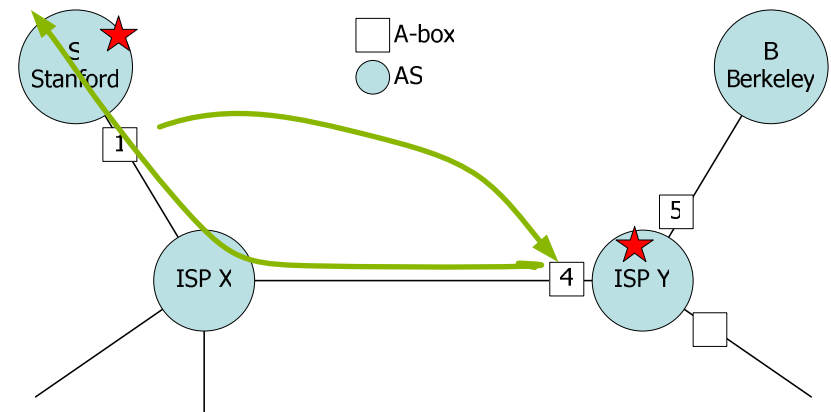
# Discovery with Partial Deployment

- I want to announce myself to every other AS
- Idea: split all ASes among all local A-Boxes, based on who is on the egress towards that AS
  - Pick a prefix per AS
  - See where I'd route it
  - If I'd route it through my AS, don't announce
  - If I'd route it outwards, send my AS number and my IP address
- If I receive an announcement
  - Respond with my own AS number and IP address
- Keep responses from ASes that are the "closest" based on the BGP tables
- Low traffic requirements
  - About 9Mbps over the **entire Internet**



# Discovery with Partial Deployment

- I want to announce myself to every other AS
- Idea: split all ASes among all local A-Boxes, based on who is on the egress towards that AS
  - Pick a prefix per AS
  - See where I'd route it
  - If I'd route it through my AS, don't announce
  - If I'd route it outwards, send my AS number and my IP address
- If I receive an announcement
  - Respond with my own AS number and IP address
- Keep responses from ASes that are the "closest" based on the BGP tables
- Low traffic requirements
  - About 9Mbps over the **entire Internet**



# Challenges

- Feasibility

- Hardware feasible (i.e., buildable). We think **YES!**
- Processing possible with current technologies
- Cost is mostly memory, much lower than high-speed routers
- Back-of-the-envelope calculations
  - 100ms reporting interval, 1 hour long-term state
  - ~400 byte avg packet size, 10 AS Internet diameter
  - 3-4.5% bandwidth overhead, 9Mbps discovery total

# Challenges (cont)

- Security
  - Intermediate hops can modify reports
  - Use (signed) iterative checks to localize lying to a single (physical) link
  - Even without security, functionality strictly better than diminishing traceroute/ping
- Incremental Deployment
  - Why would ISPs adopt ASTRA?
  - What would be the benefit for the 1st ISP?
  - Security works only for fully-deployed path prefixes

# Next Steps

- Evaluation
  - Trace-based validation of back-of-the-envelope calculations
- Generalization of audit architecture
  - Traceback can use same machinery
  - Track latencies (simple HW extension)
  - An AS-level traceroute tool (trivial HW extension)
  - An AS-level loss map of my egress spanning tree of the Internet
  - What else can an A-box do for you?
- Refinements
  - Packet transformations
  - Multicast
  - Report cascades
- Liar disambiguation based on scoped broadcast dispute resolution
- Proof-of-concept hardware spec

# What's the big deal?

- Current practice (pings, traceroutes) operates at the whim of ISPs
  - Works with probe traffic only
  - Exposes intra-AS info which ISPs want to keep private
- Our philosophy: let ASes be opaque
  - Get inter-AS information only
  - Keep ASes in the loop; they have custody of the monitoring infrastructure
- Incentives?
  - Good ISPs will want to use this: shows their neighbors that they're good
  - Market pressure could (should) convince the rest
  - Famous last words? Perhaps...
- Where did QoS go?
  - If I can find the good paths, I don't need to reserve any
- Did you break my layering?
  - Nope. IP semantics still the same
  - We just provide an external data source for the Knowledge Plane

# Q&A

- Thank you