

Avoiding Cascading Failures



SRECON2016 - Craig Fender - Ravindra Punati



Avoiding Cascading Failures



Craig Fender

Craig is presently a Senior Technical Duty Officer at eBay and is responsible for commanding all types of large scale site incidents. In addition to an undergraduate degree Craig holds numerous professional certifications related to the computer industry (RHCE, SSCE, ITIL and et cetera). Craig has held several roles at multiple start-up and fortune 500 companies such as Senior Systems Engineer, Project Manager, Presenter and Major Incident Commander.

Ravindra Punati

Ravindra Punati is leader of the Site Reliability Engineering team. In other roles at eBay Ravi has been responsible for the infrastructure automation initiatives and cloud operations. Ravi brings extensive expertise in the fields of database engineering, application development and software as a service product lines. In addition to holding multiple degrees in computer science Ravi has held several roles as an engineer, architect, manager and executive in various silicon valley start-ups.

EBAY AT A GLANCE

OUR BUSINESS



EBAY AT A GLANCE



\$8.6B

Revenue in 2015



162M

Global Active Buyers *



800M

Live Listings *



57%

Percentage of eBay Inc.
revenue that is
international *



190

Markets eBay apps
are available in*



\$82B

GMV in 2015

*Q4 2015 data

EBAY LEADS IN MOBILE

304M

eBay Application
downloads *

43%

Percentage of GMV
closed on Mobile *

\$33B

2015 MCV
across eBay's
portfolio of apps

Every

28 sec.

a tablet is bought via
mobile in the U.S. *

Every

10 sec.

a ladies handbag is
bought via mobile in
the U.S. *

9.2M

new listings are
added via mobile
every week *

*Q4 2015 data

EBAY OPERATIONS

Background on eBay Operations



Background on eBay Operations



Background on eBay Operations

- In any mission critical real-time environment where one is highly incentivized toward maximum uptime one must have:
 - redundancy ,
 - resiliency and
 - robustness.
- These needs give rise to complexity.
- Complexity can lead to fragility.
 - Multilayered dependencies can cause cascading failure.
- To manage that complexity certain behaviors, culture, principles and technology emerge as the most successful.

Background on eBay Operations

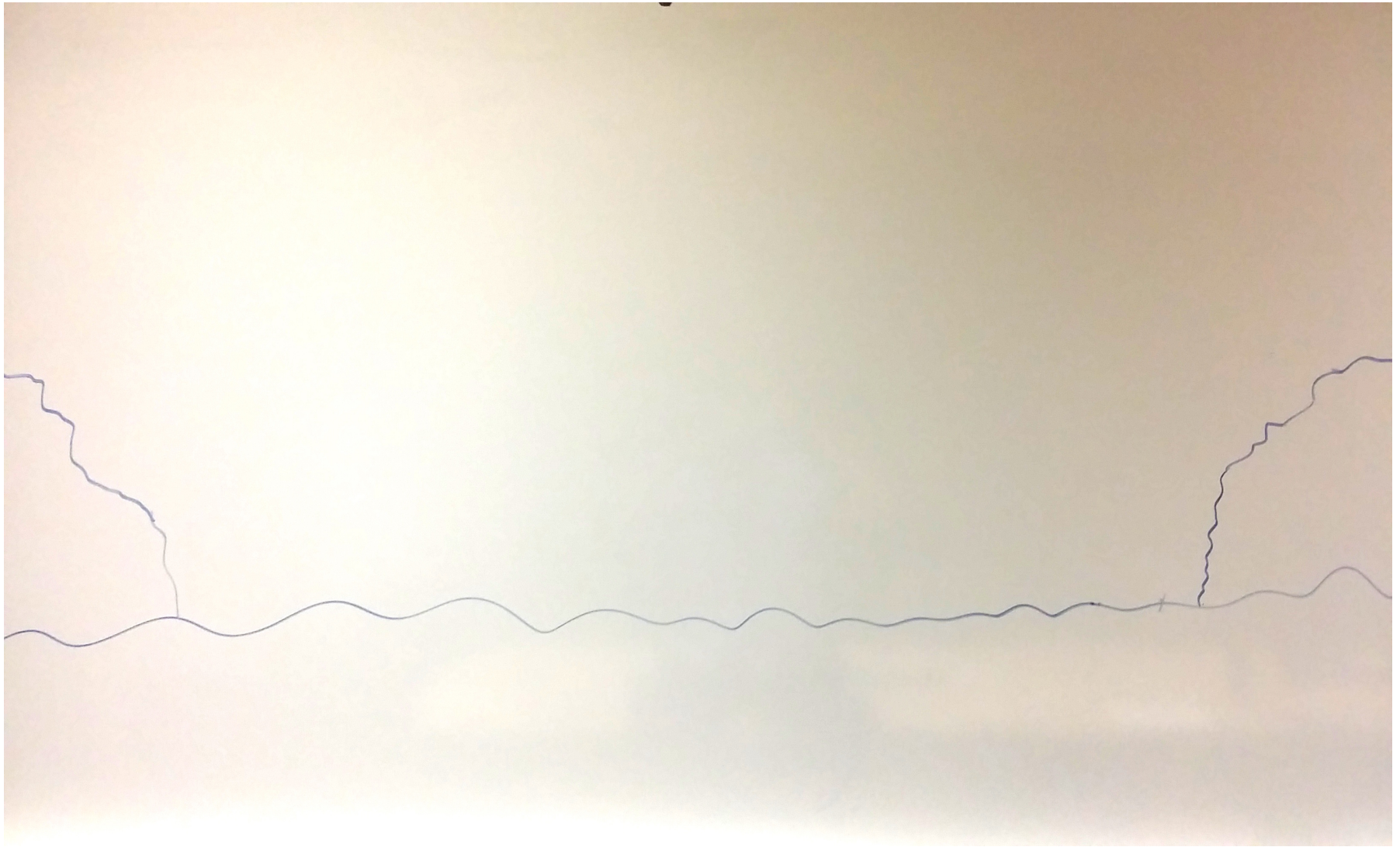
- eBay itself is like a bridge between buyers . . .



Background on eBay Operations

- . . .and sellers

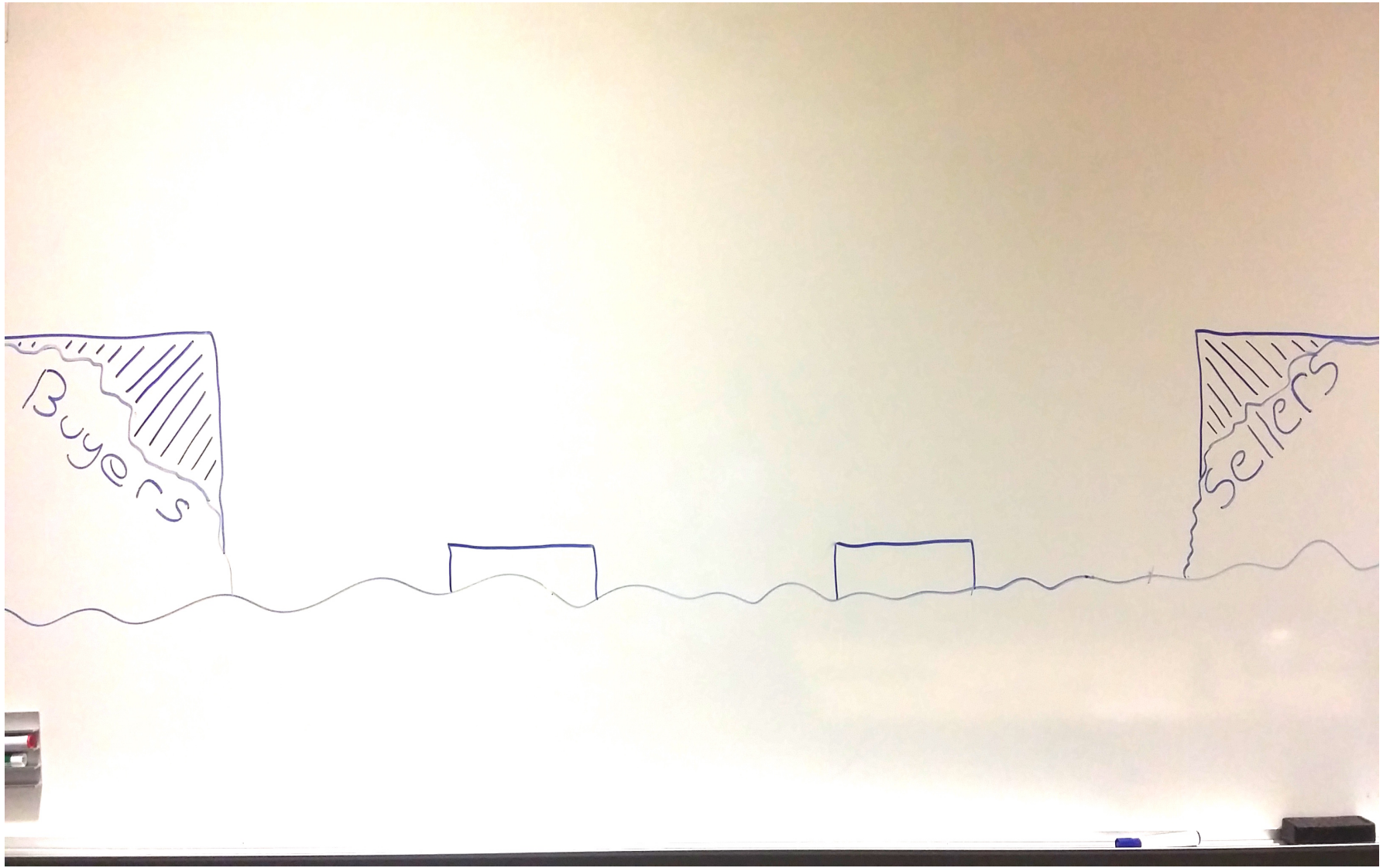


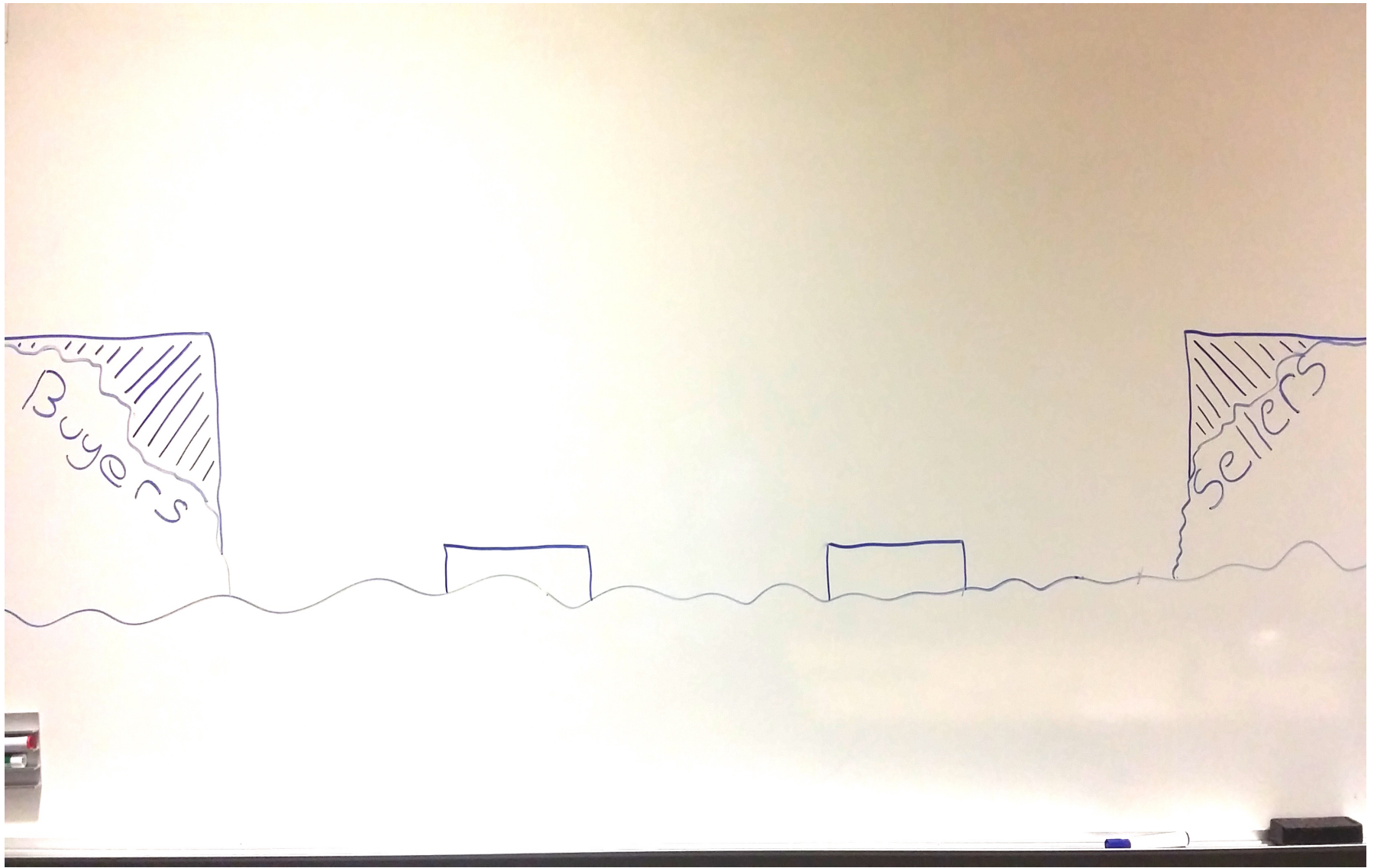


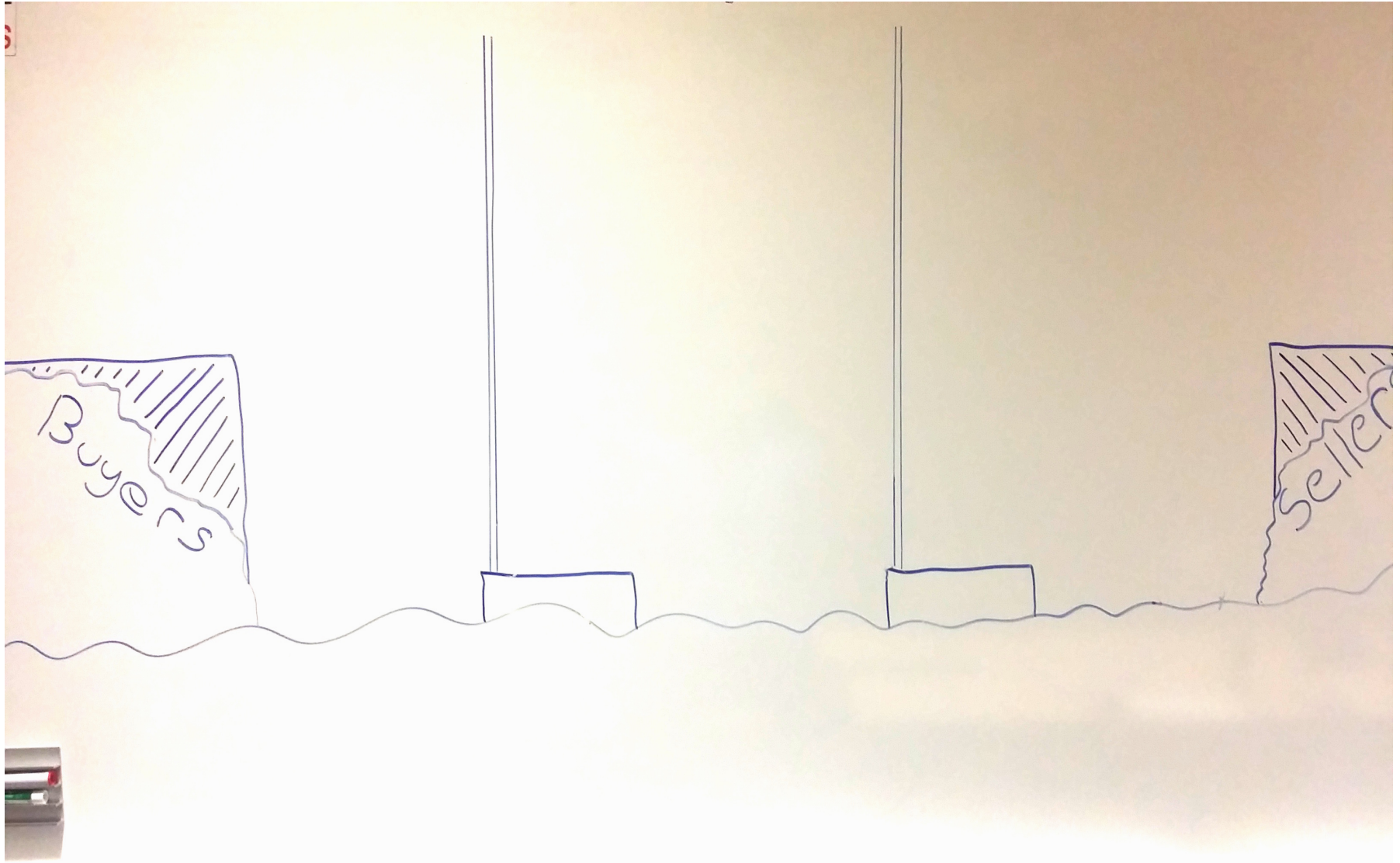


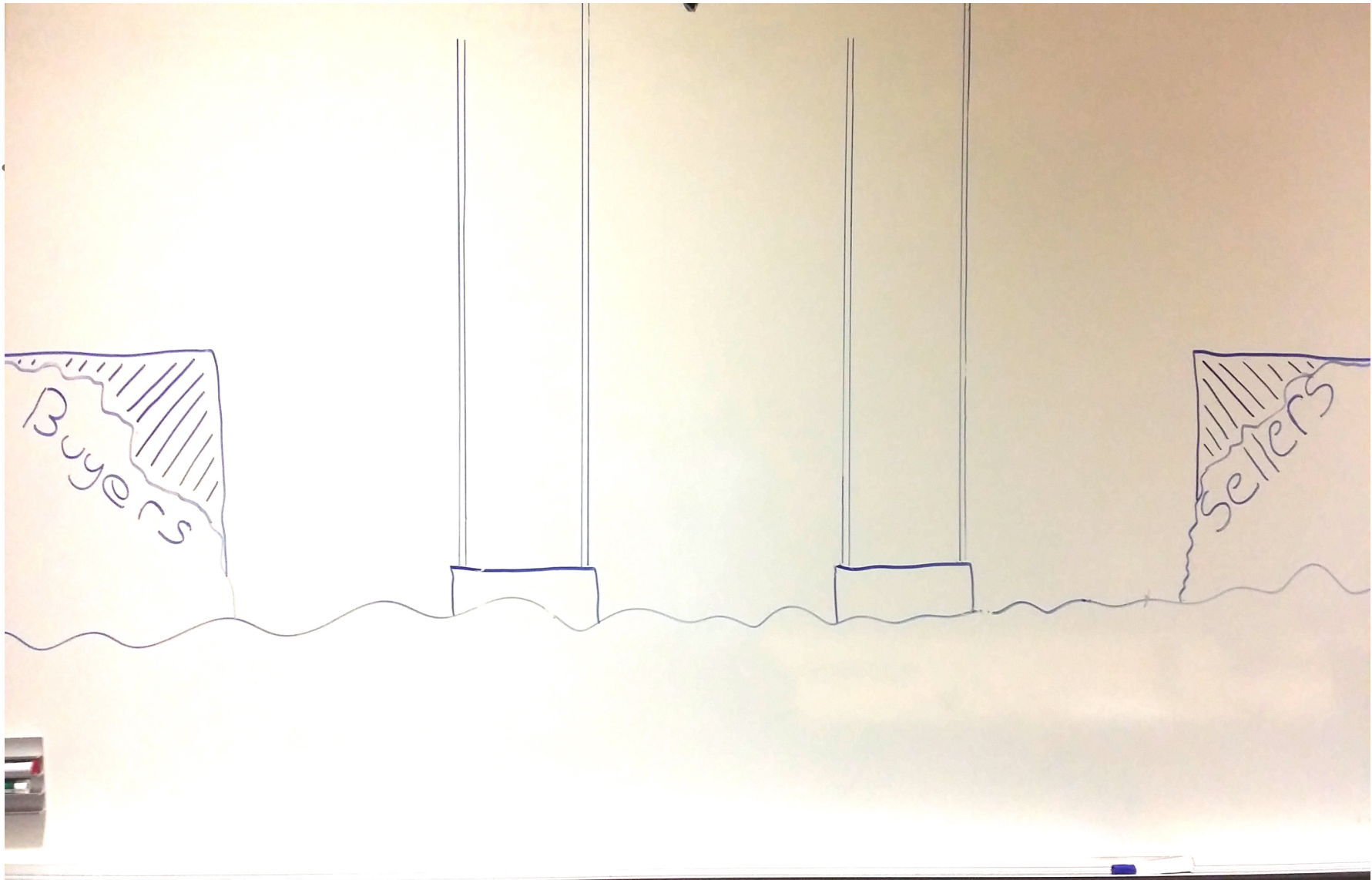
Buyers

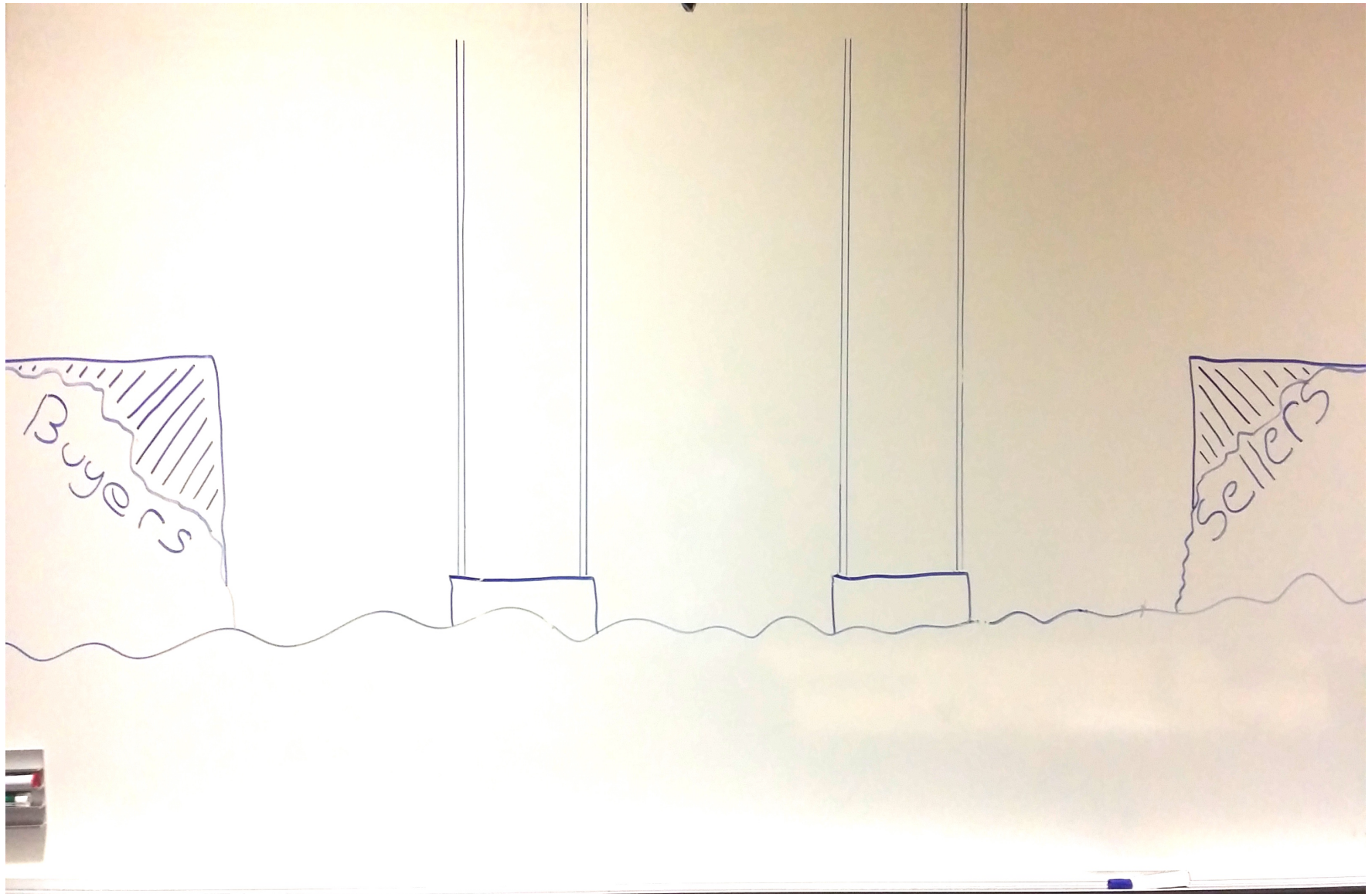
Sellers

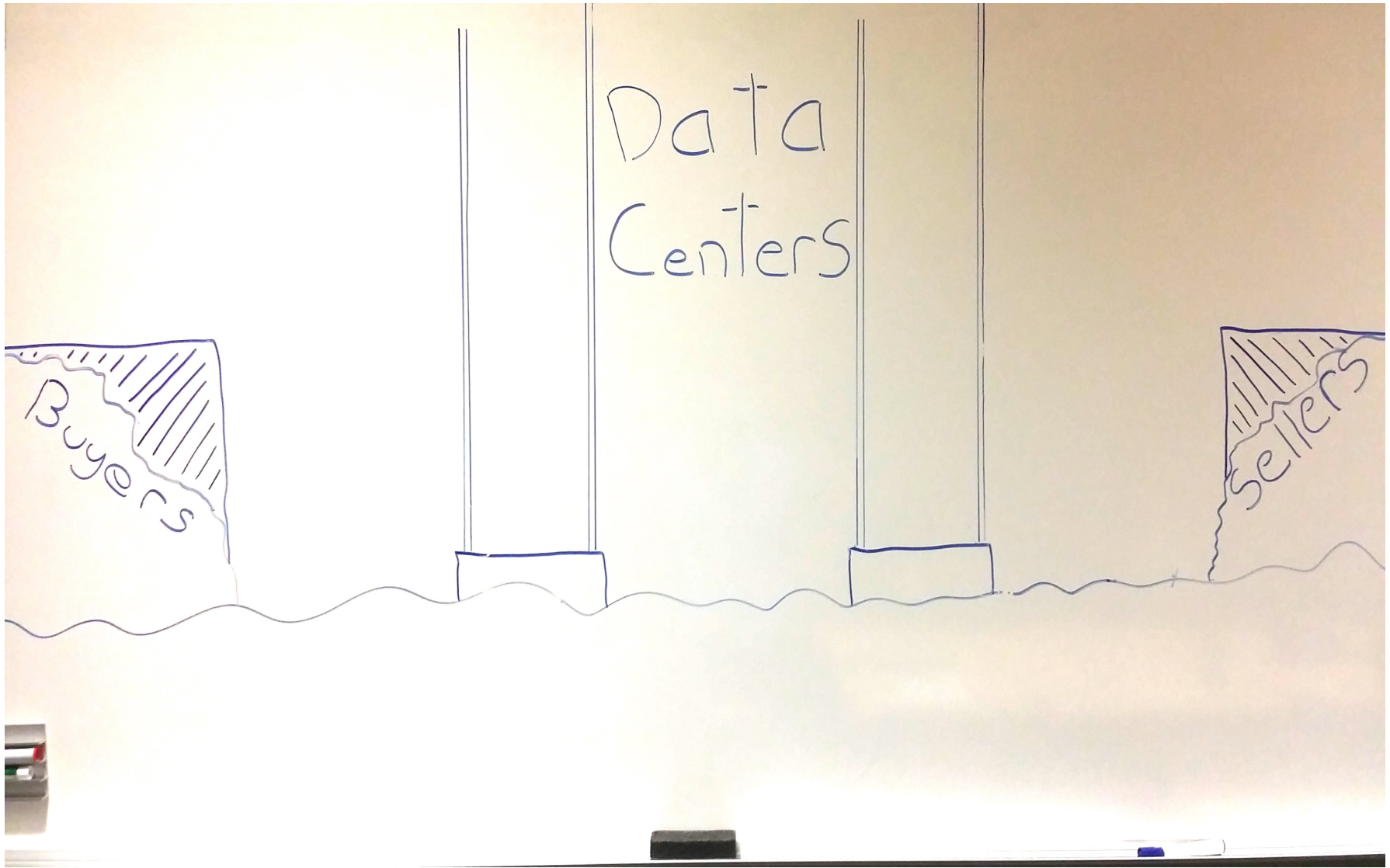


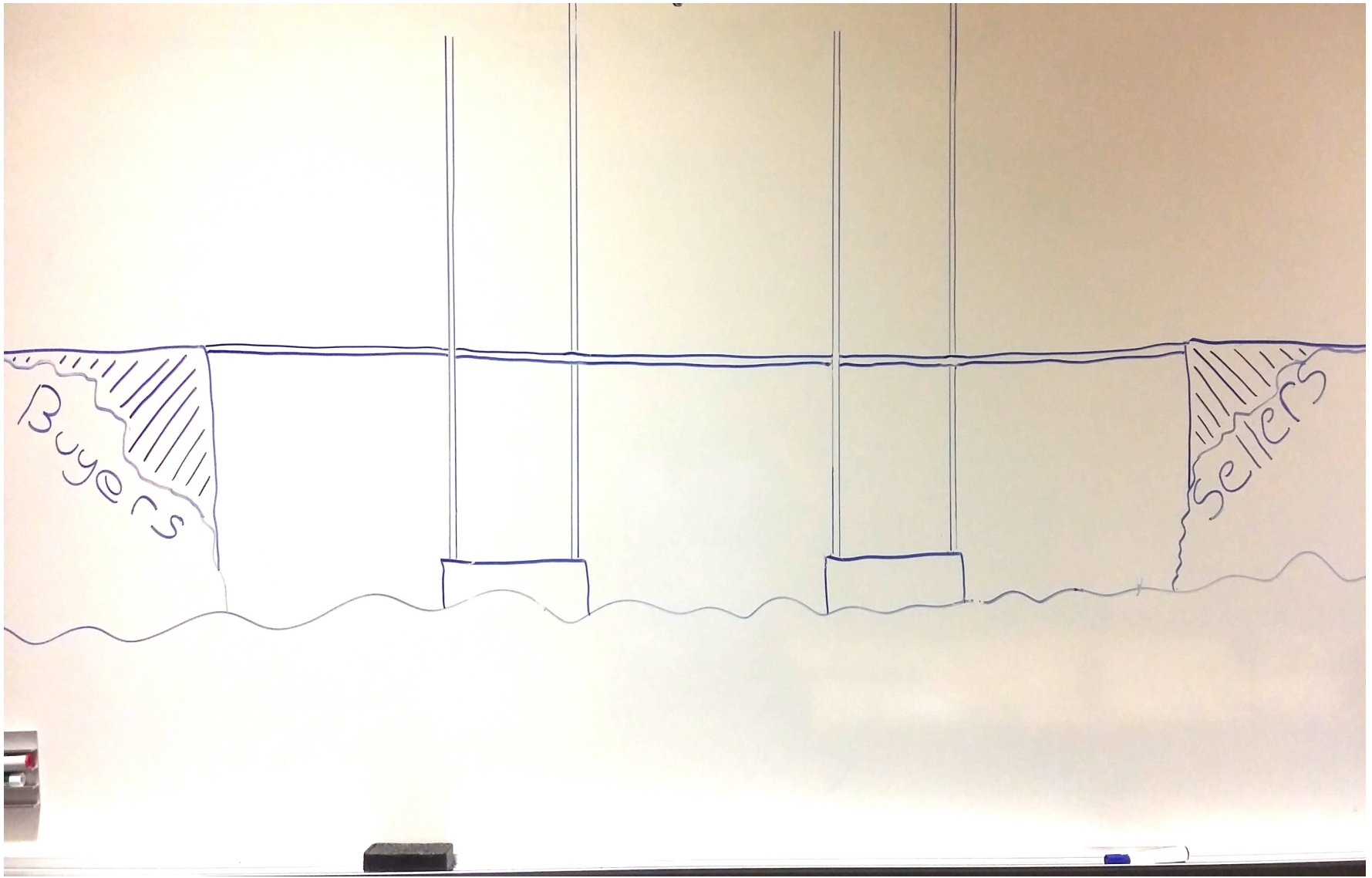


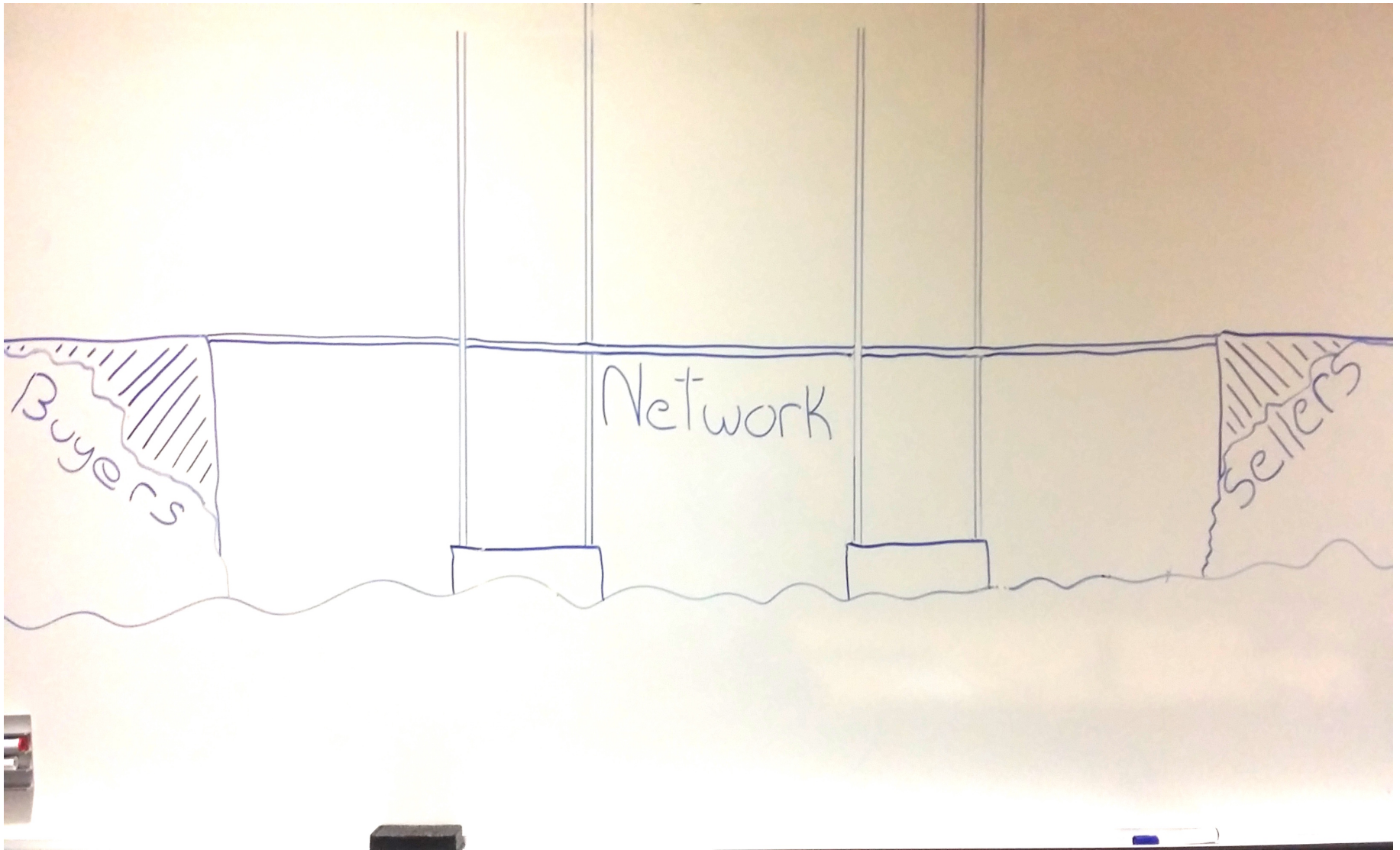


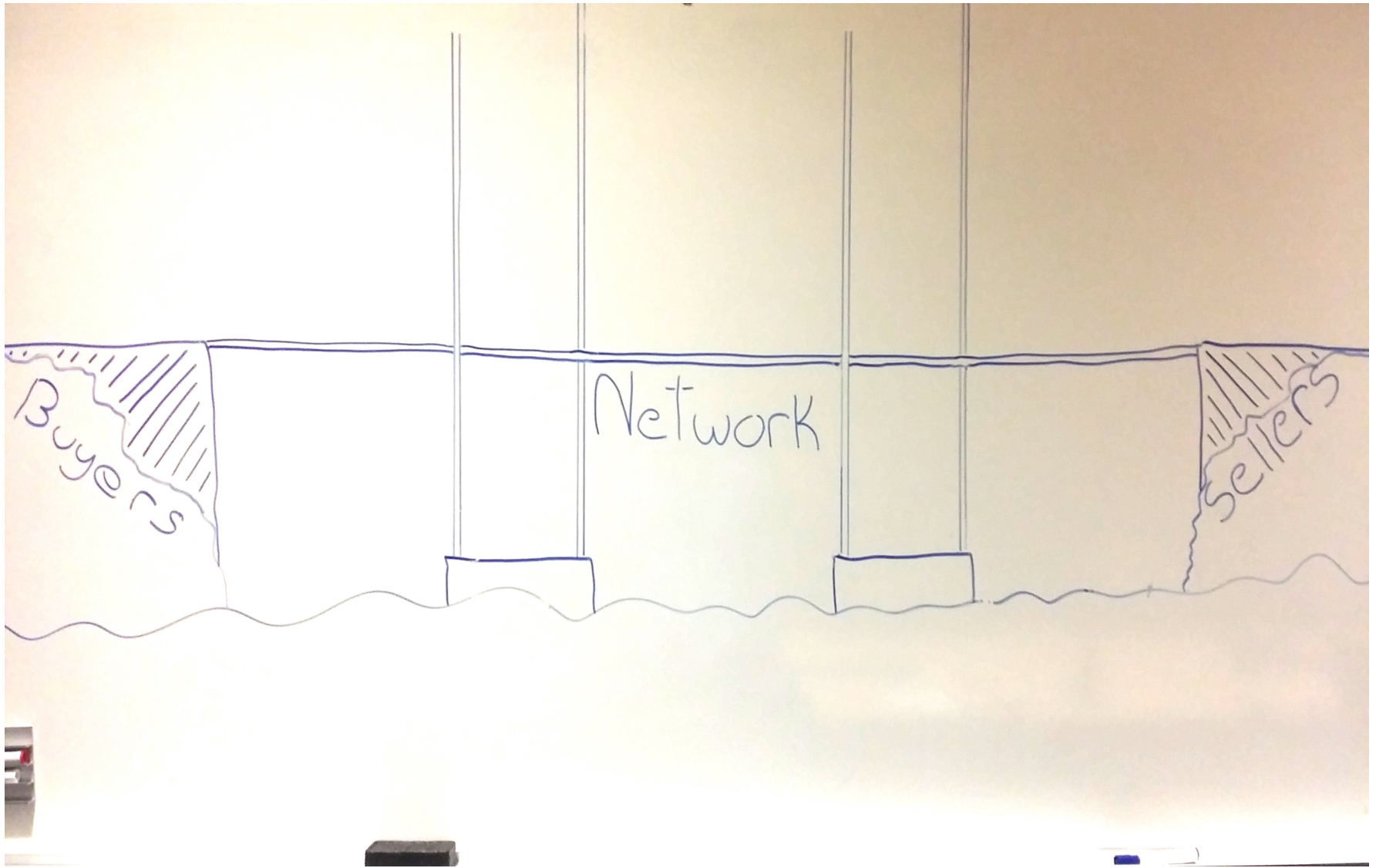


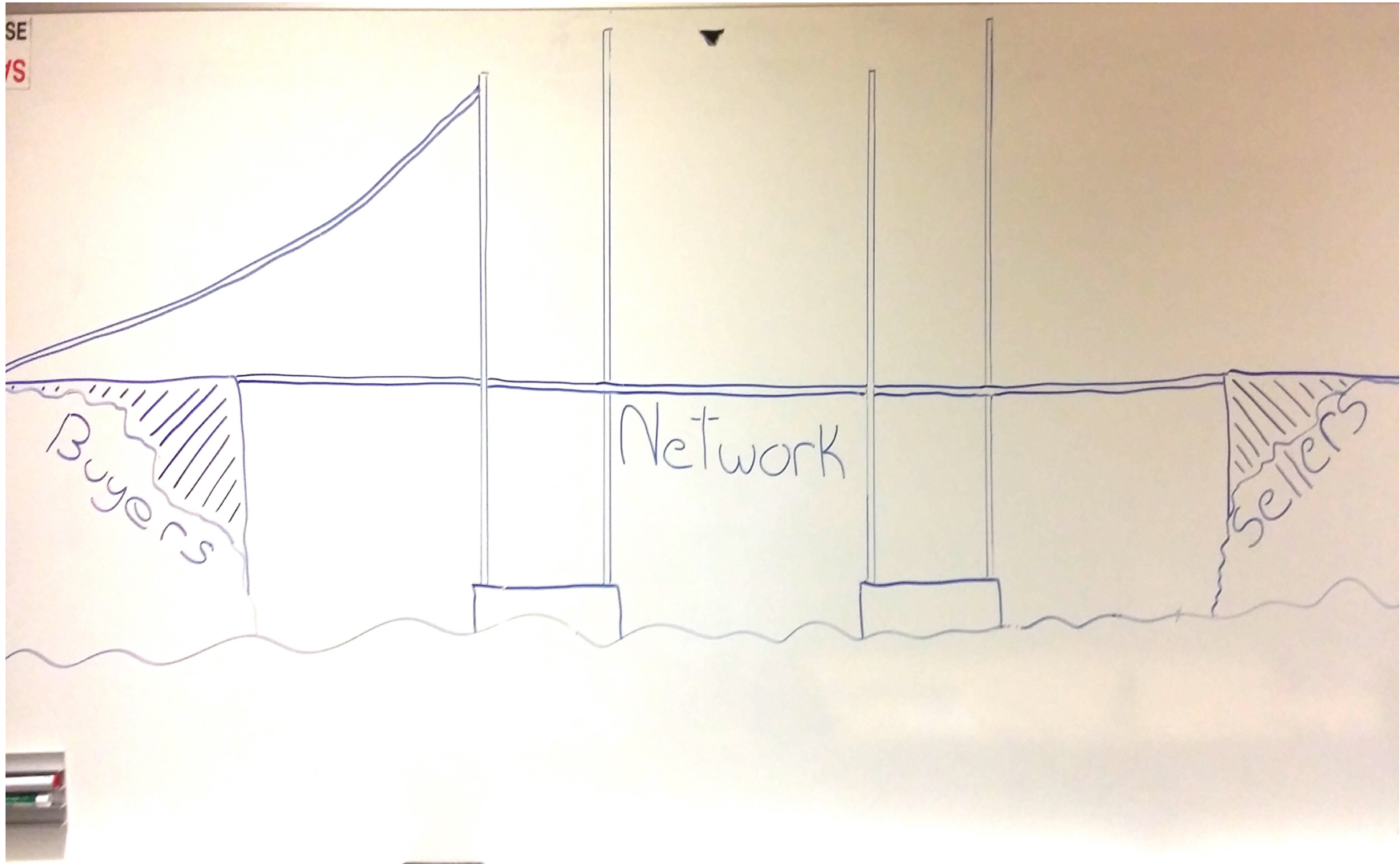


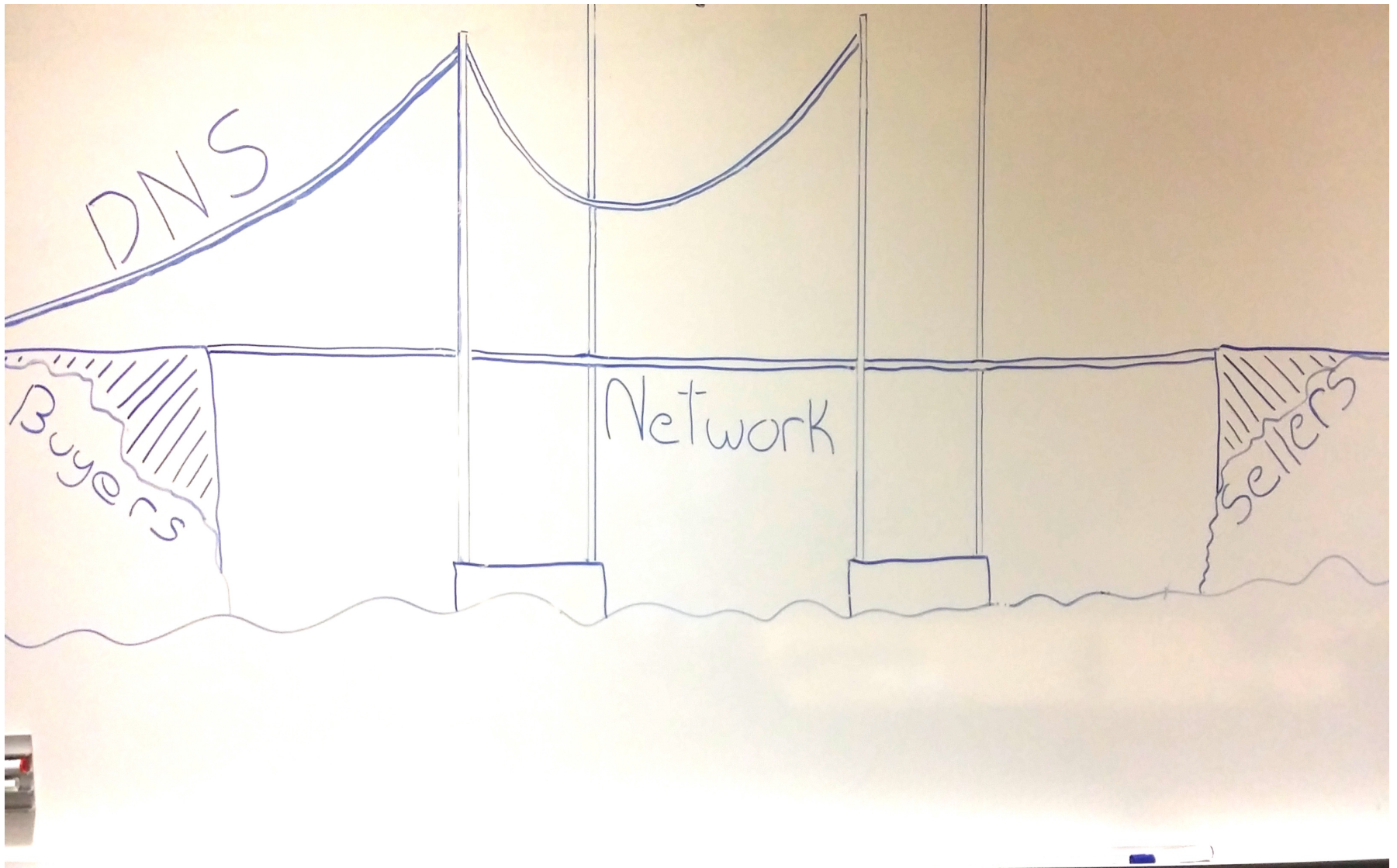


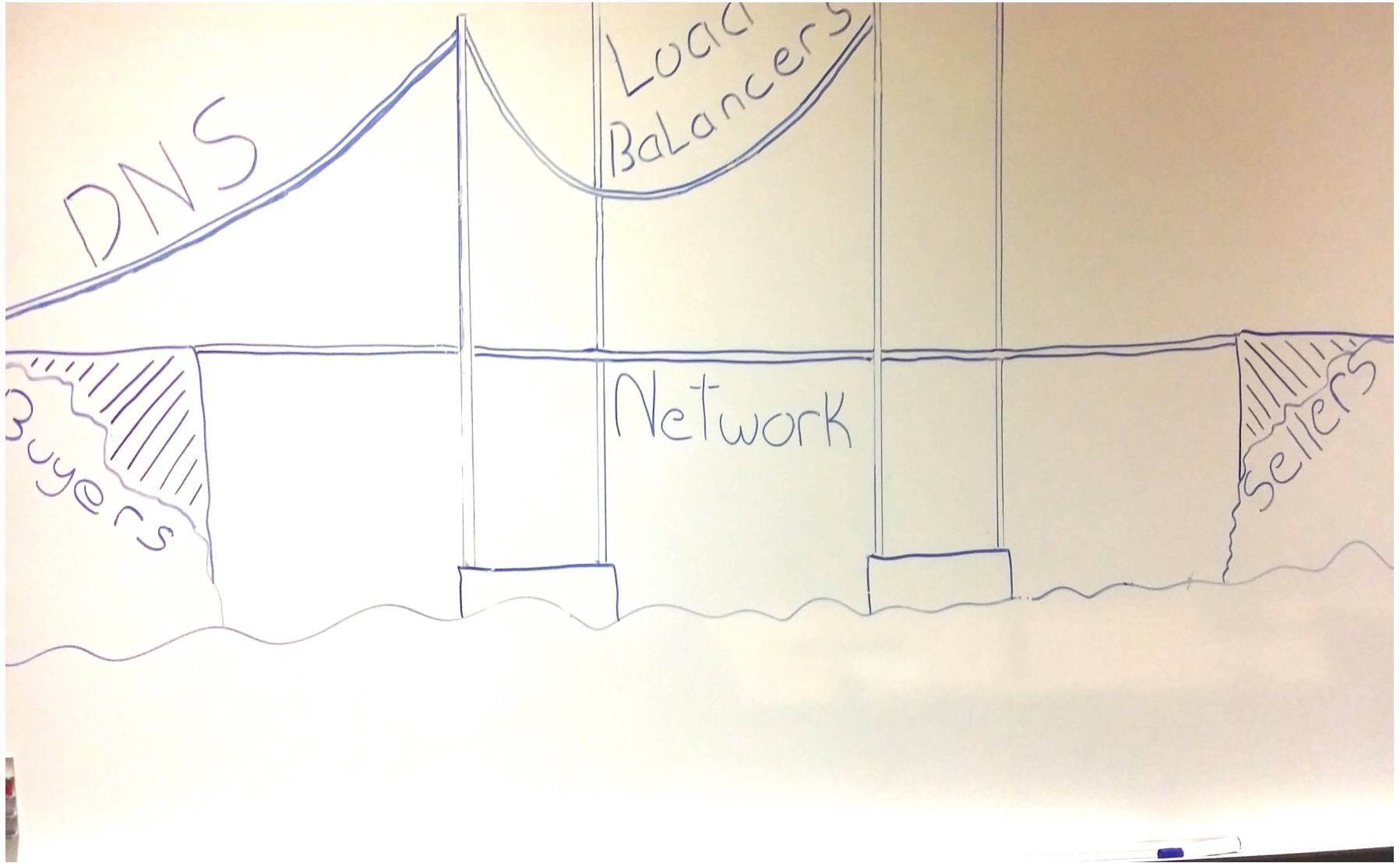


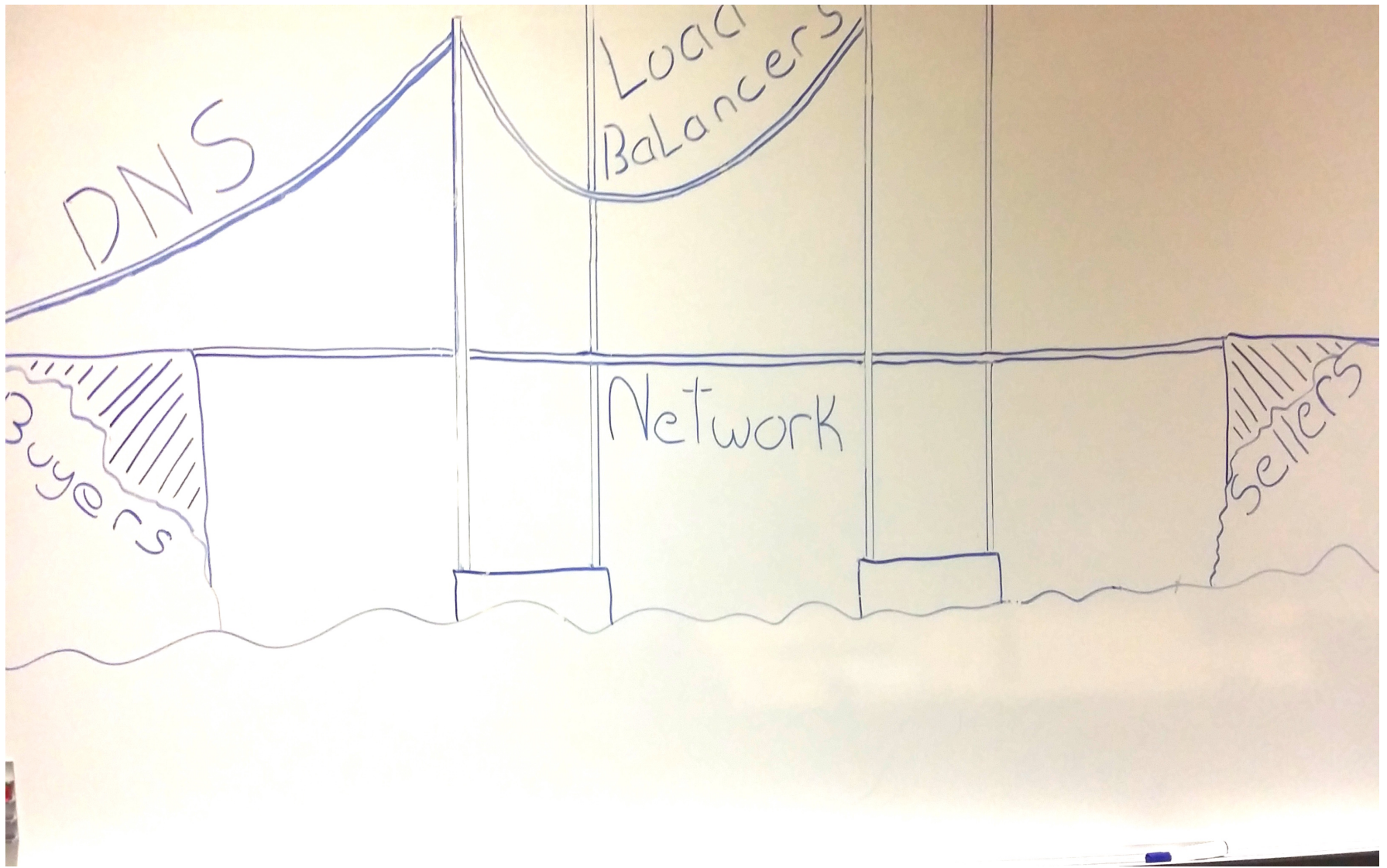


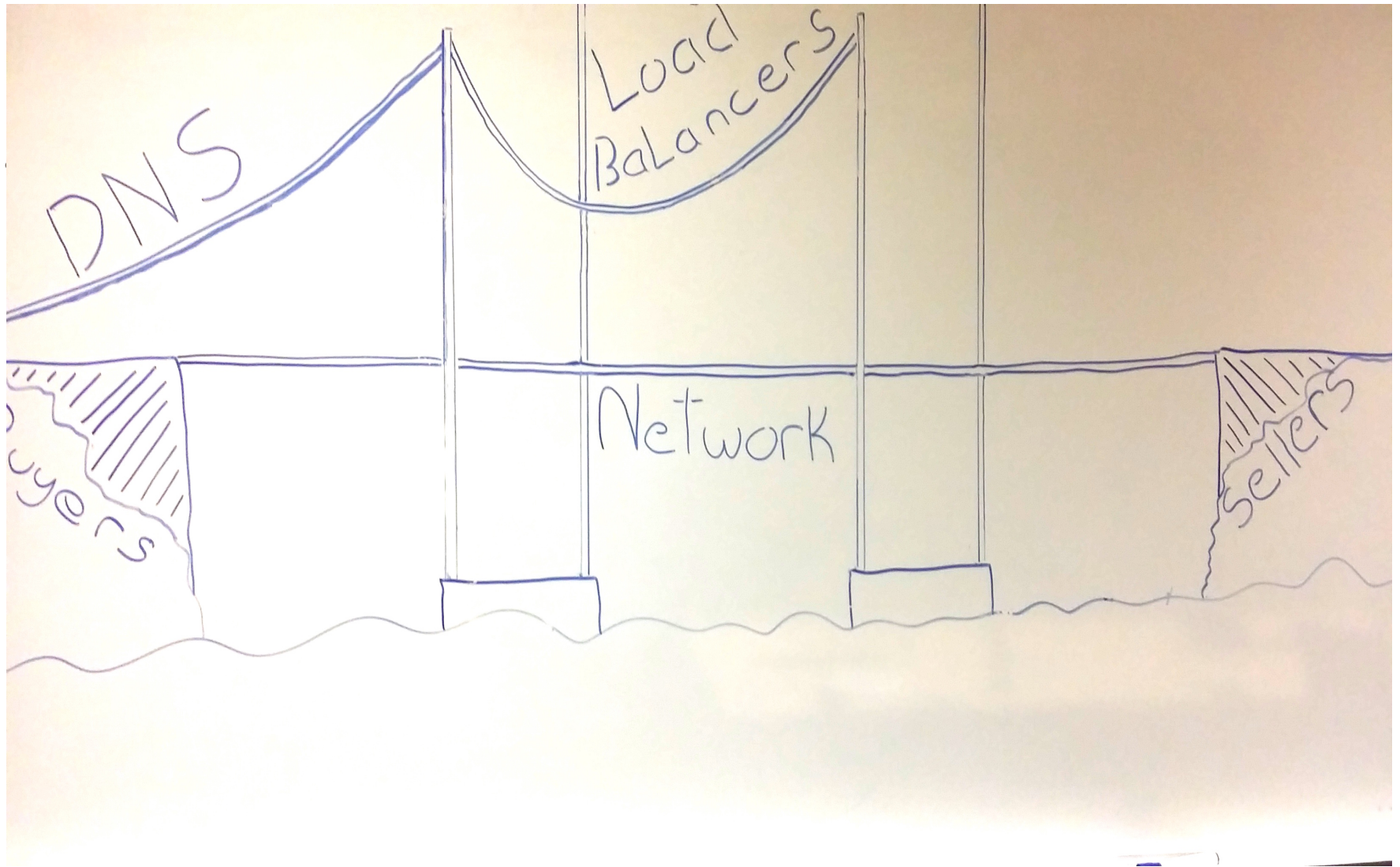


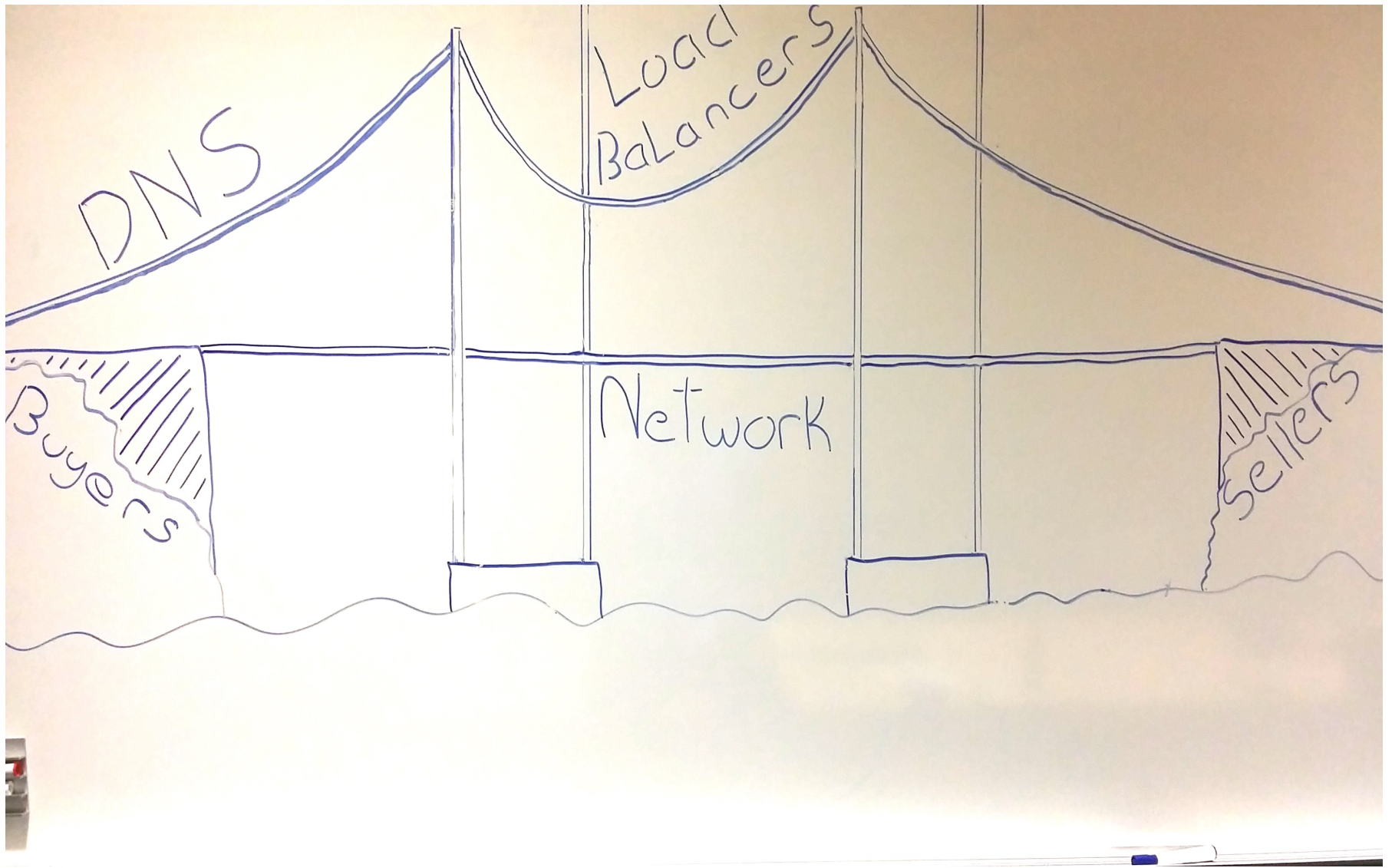


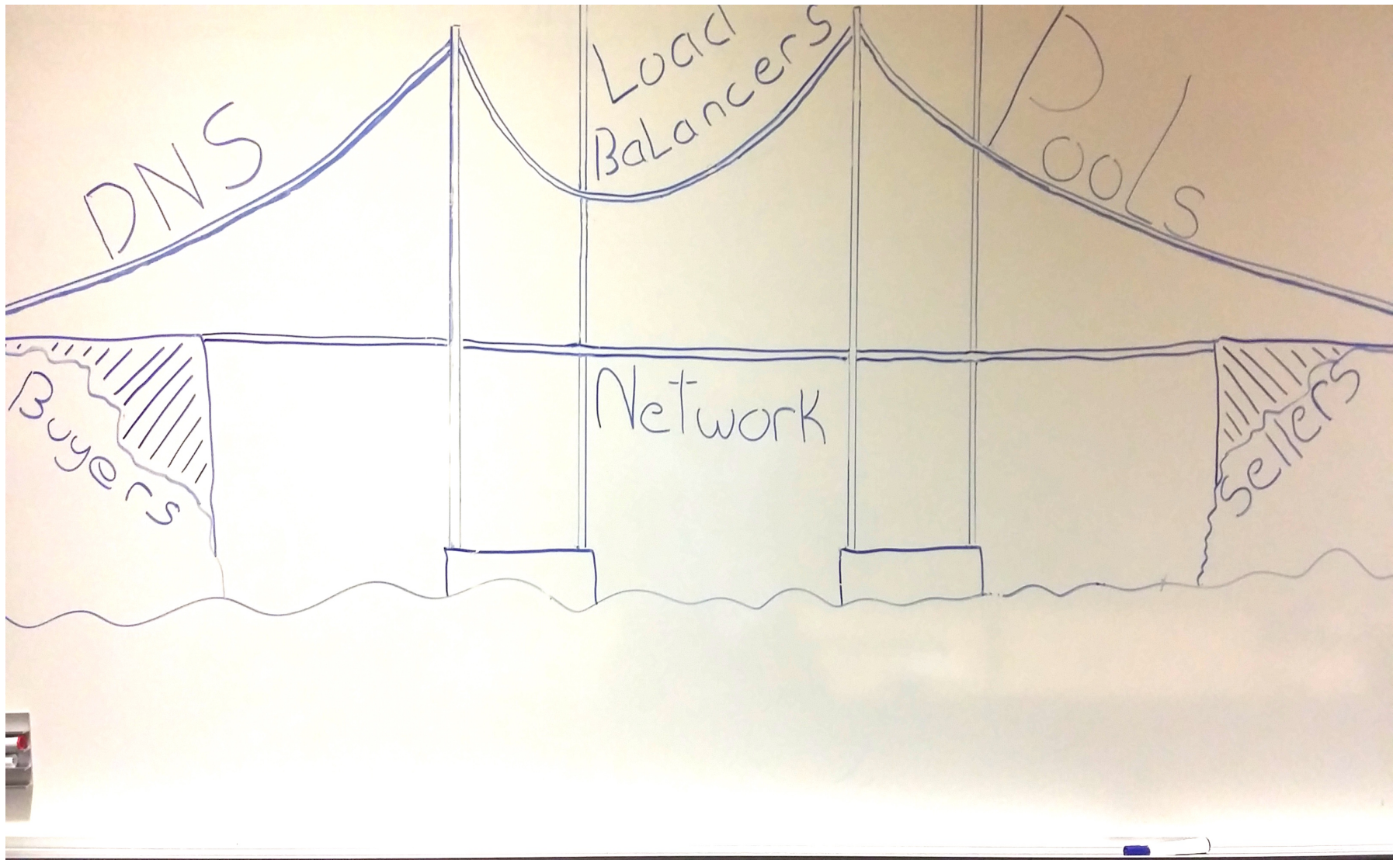


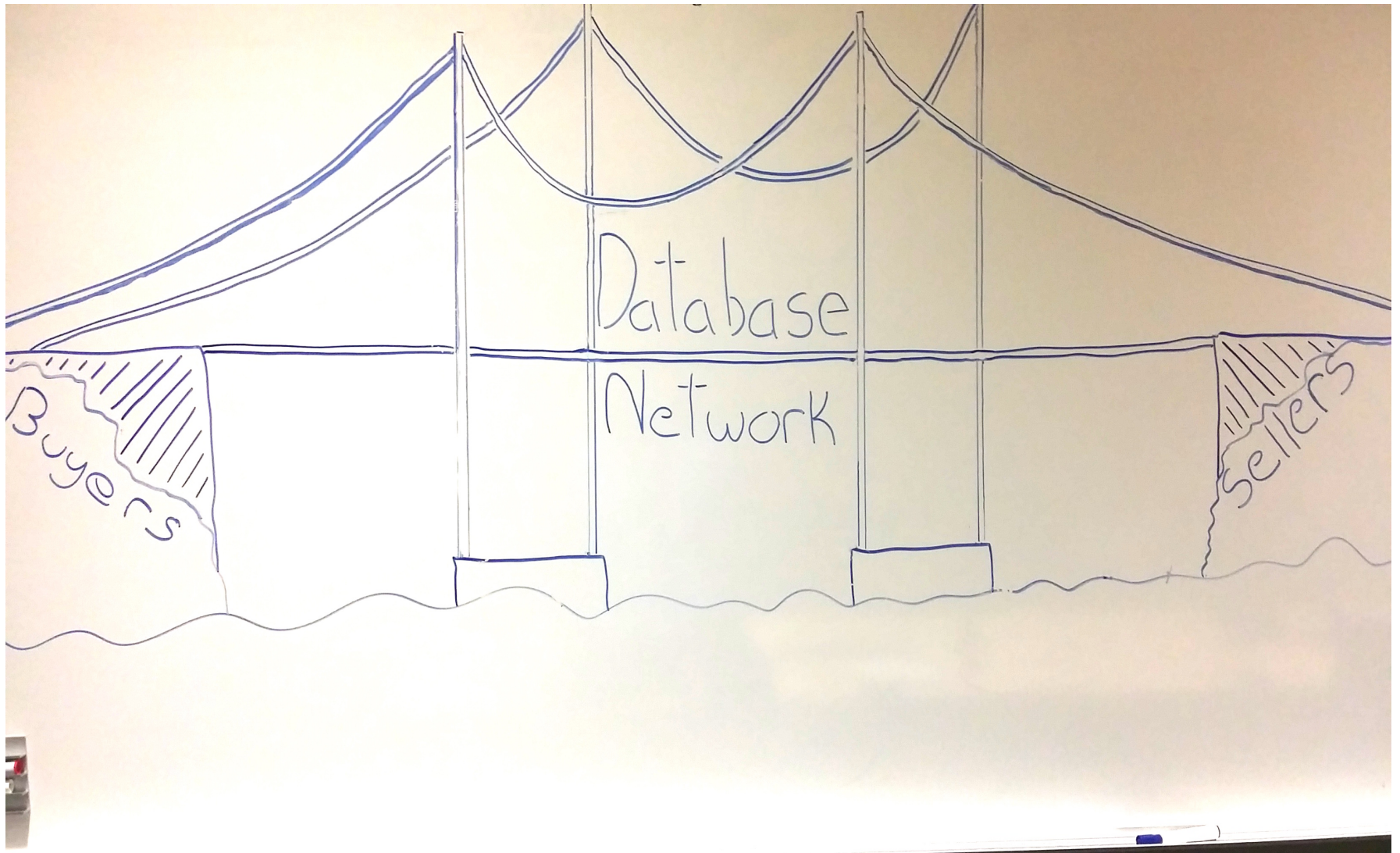


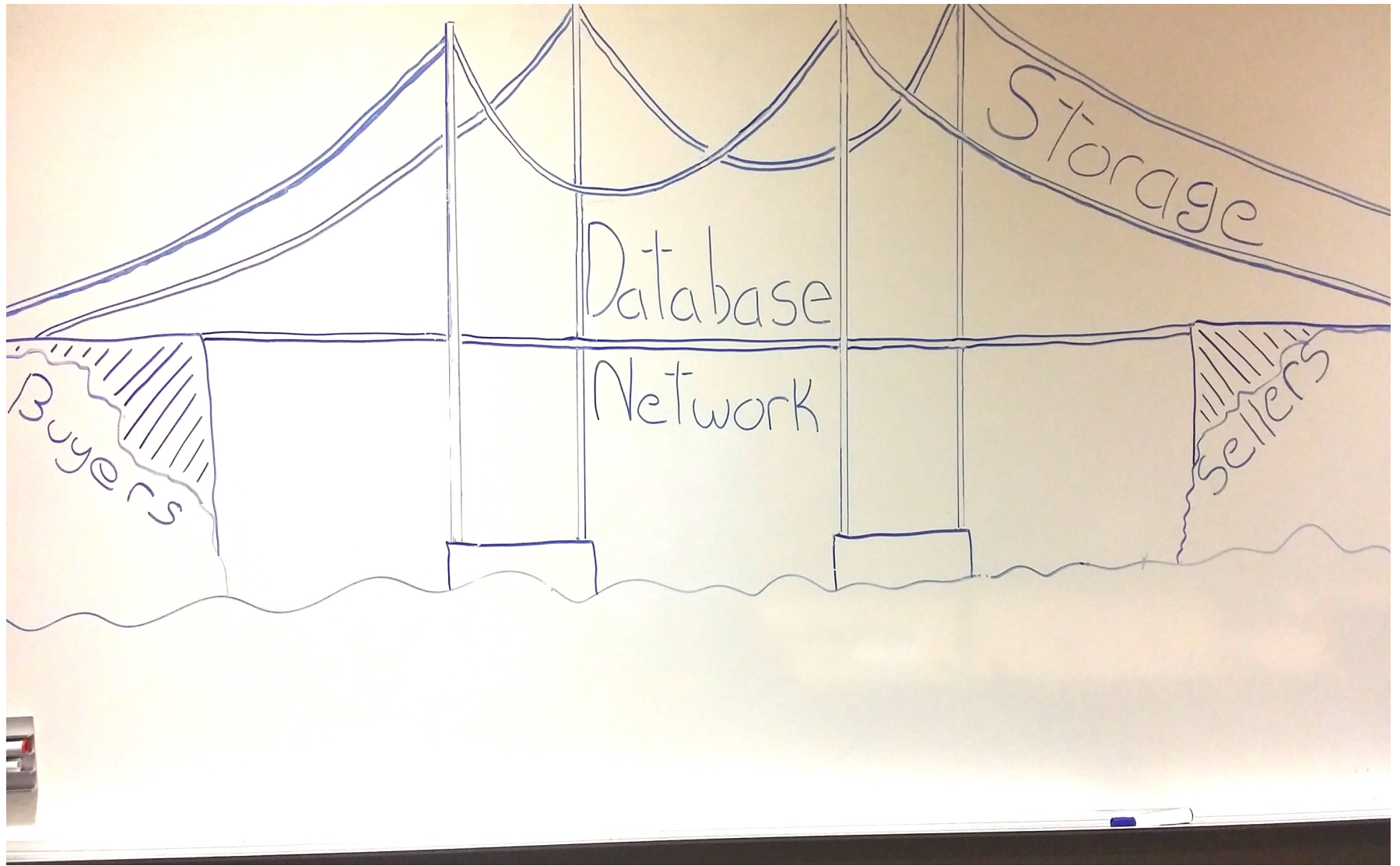


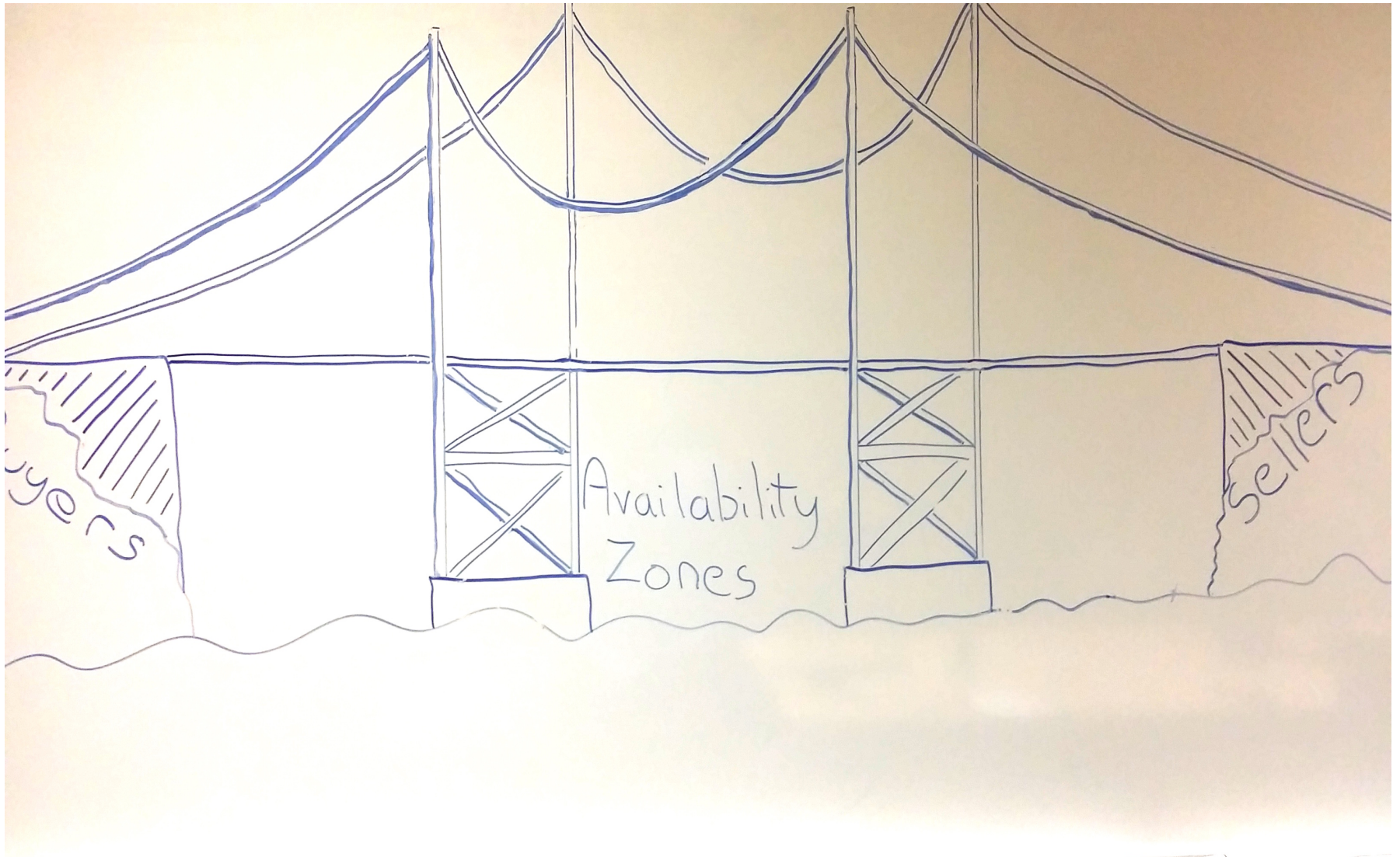


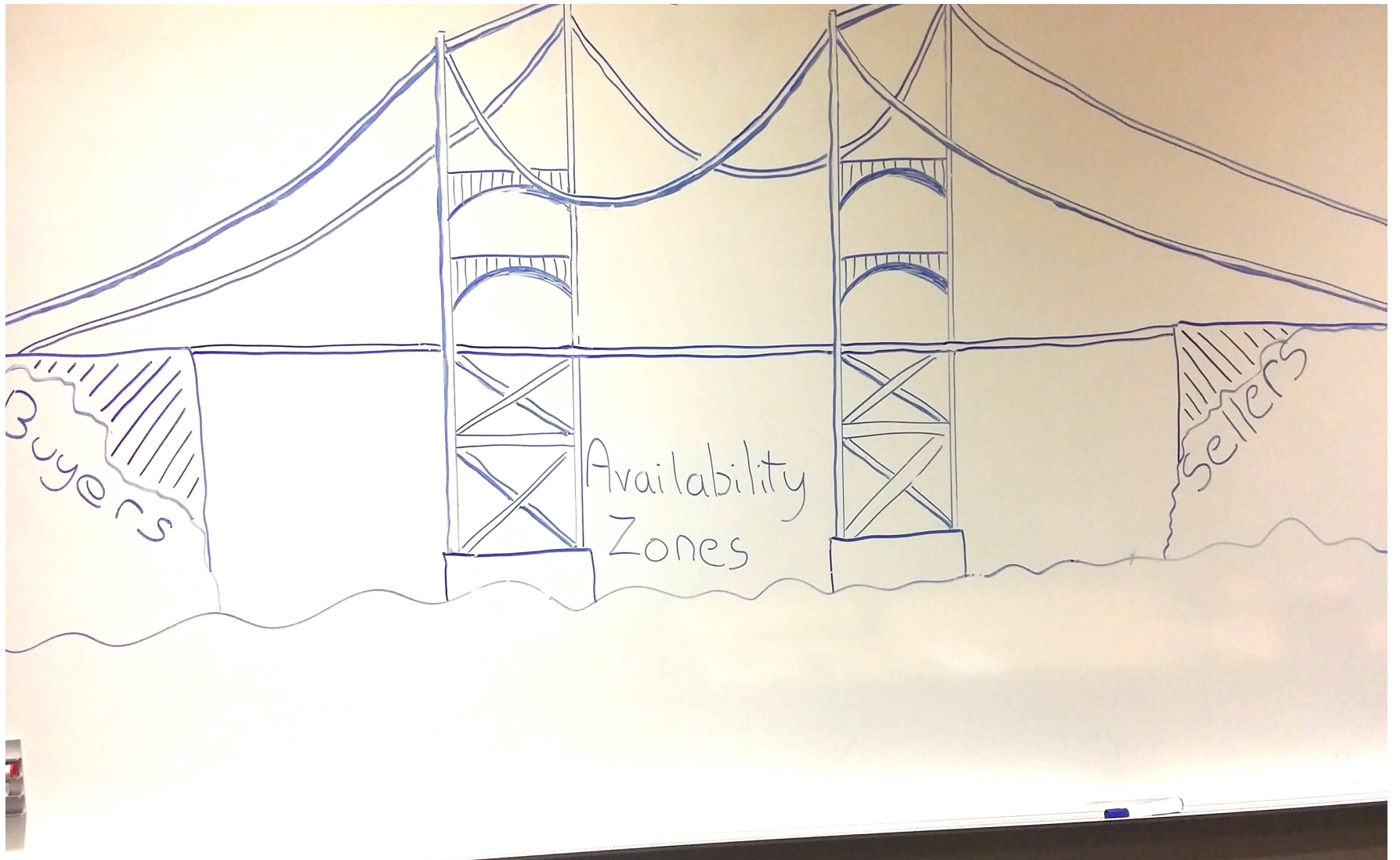


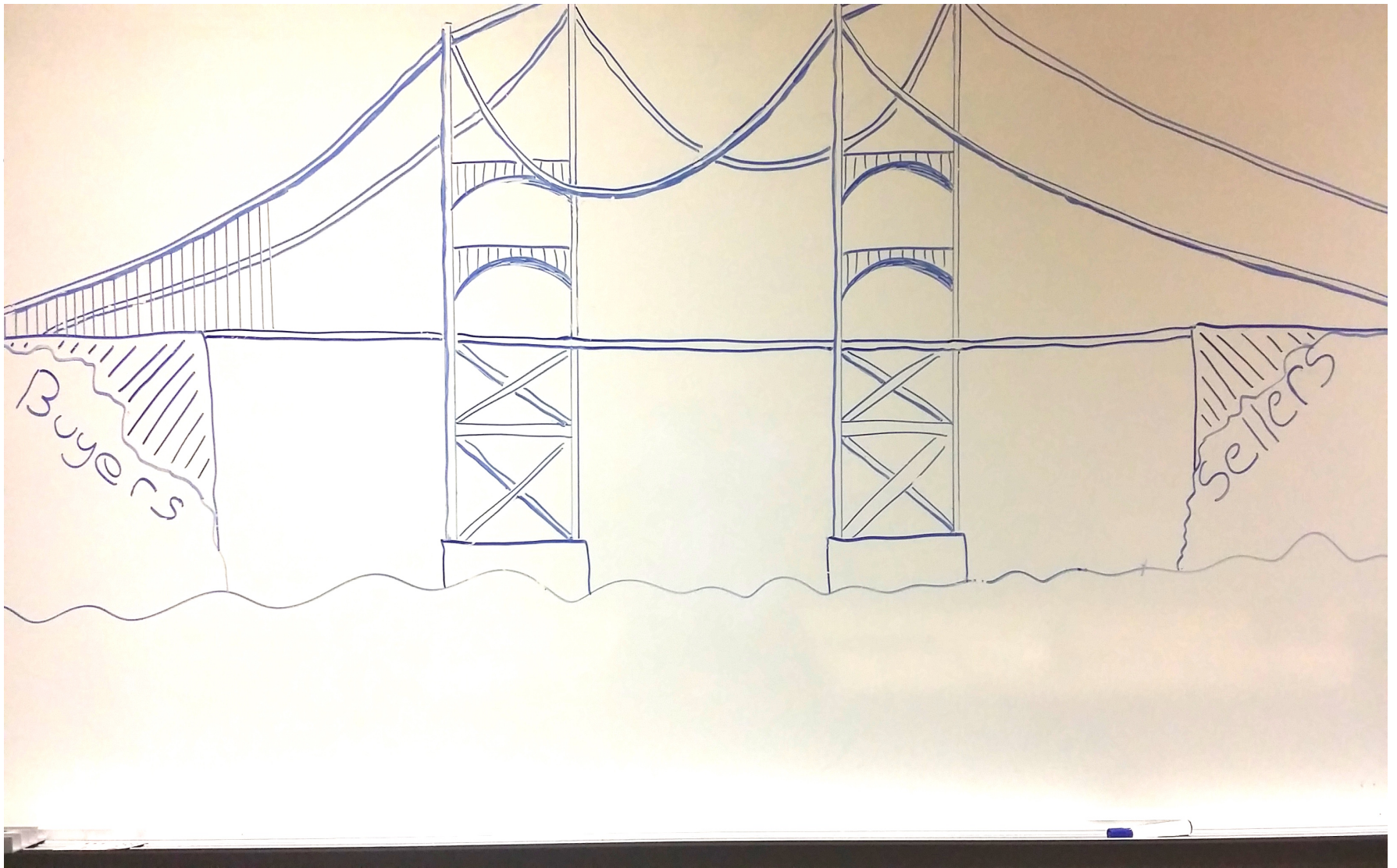


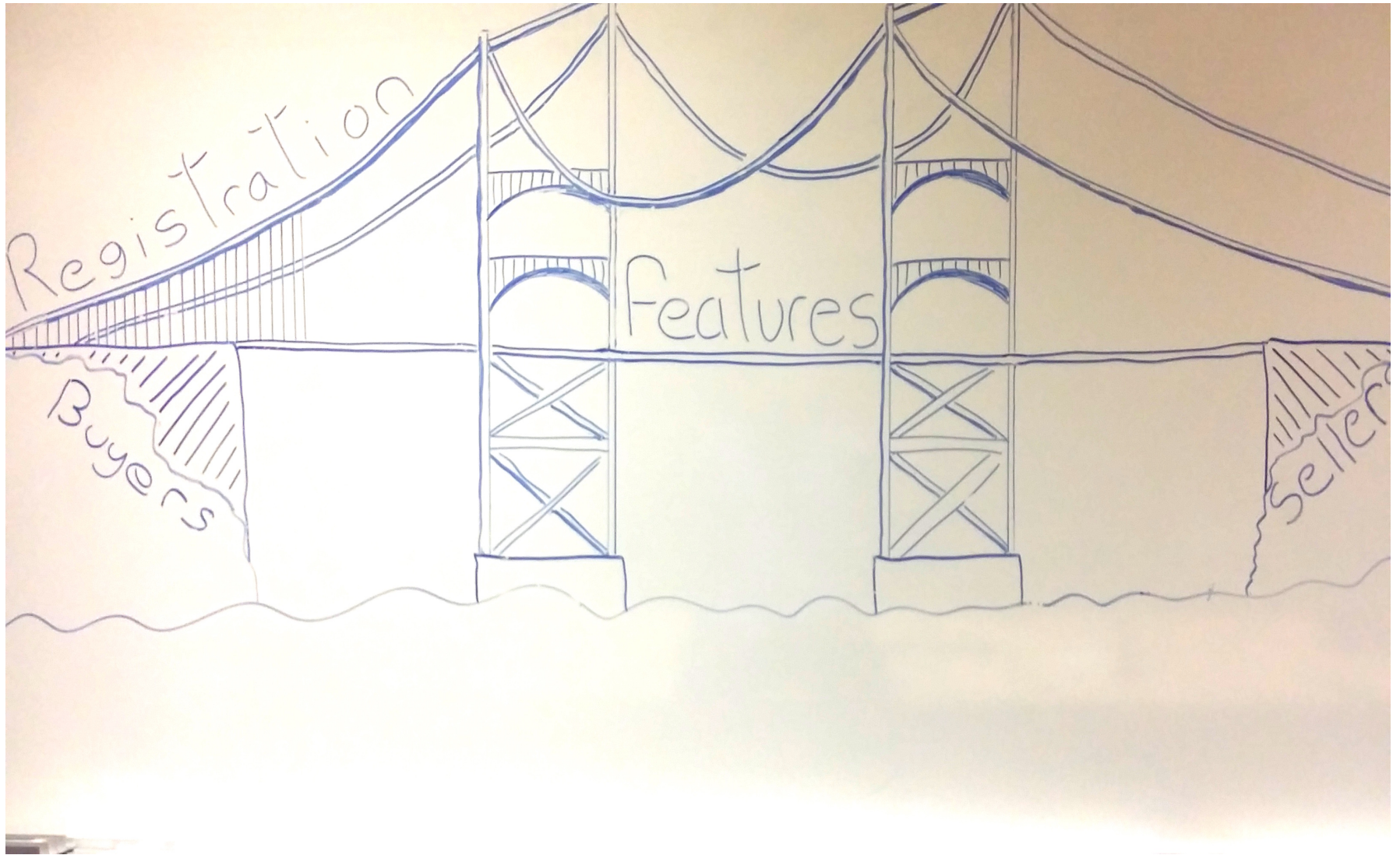


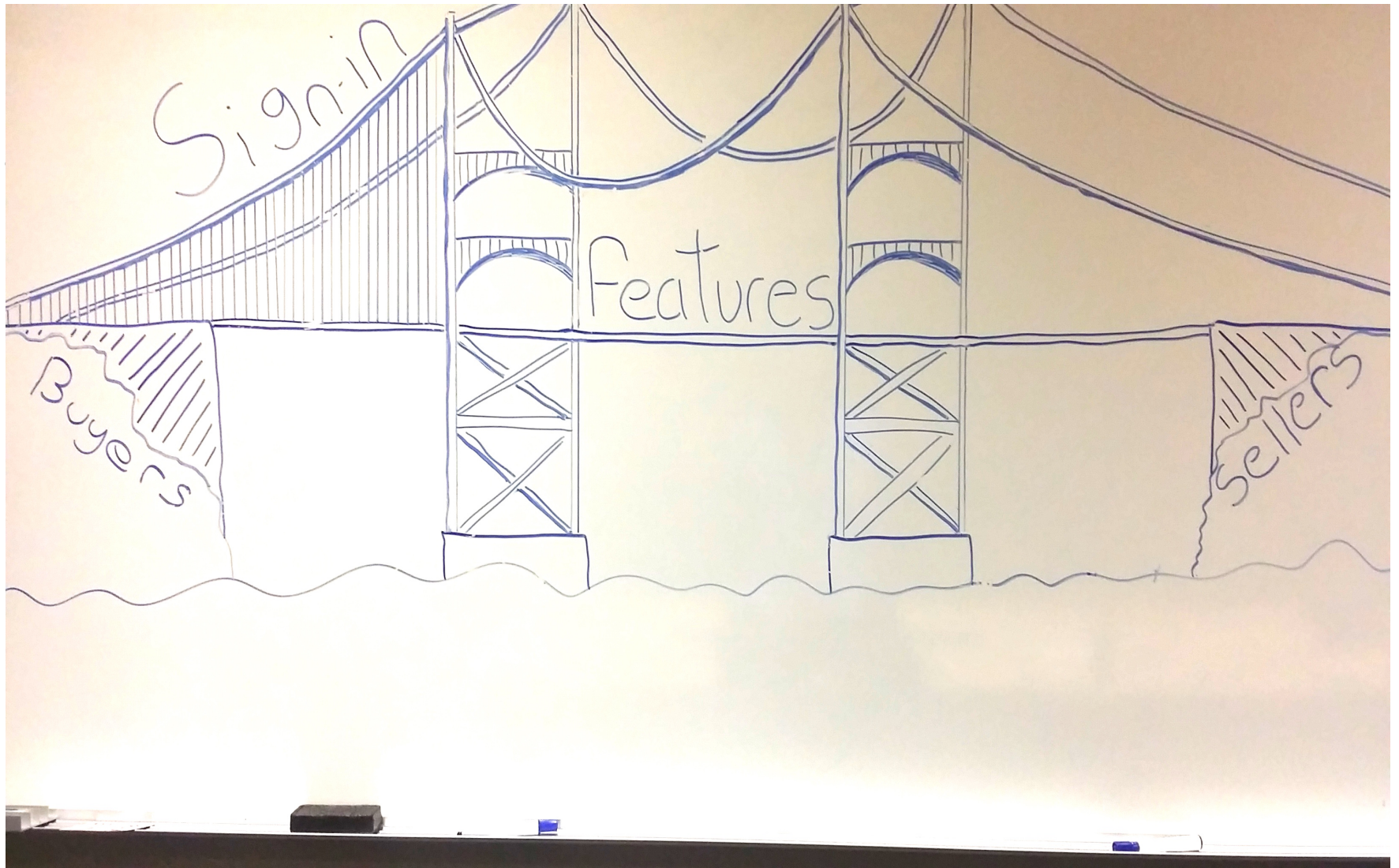


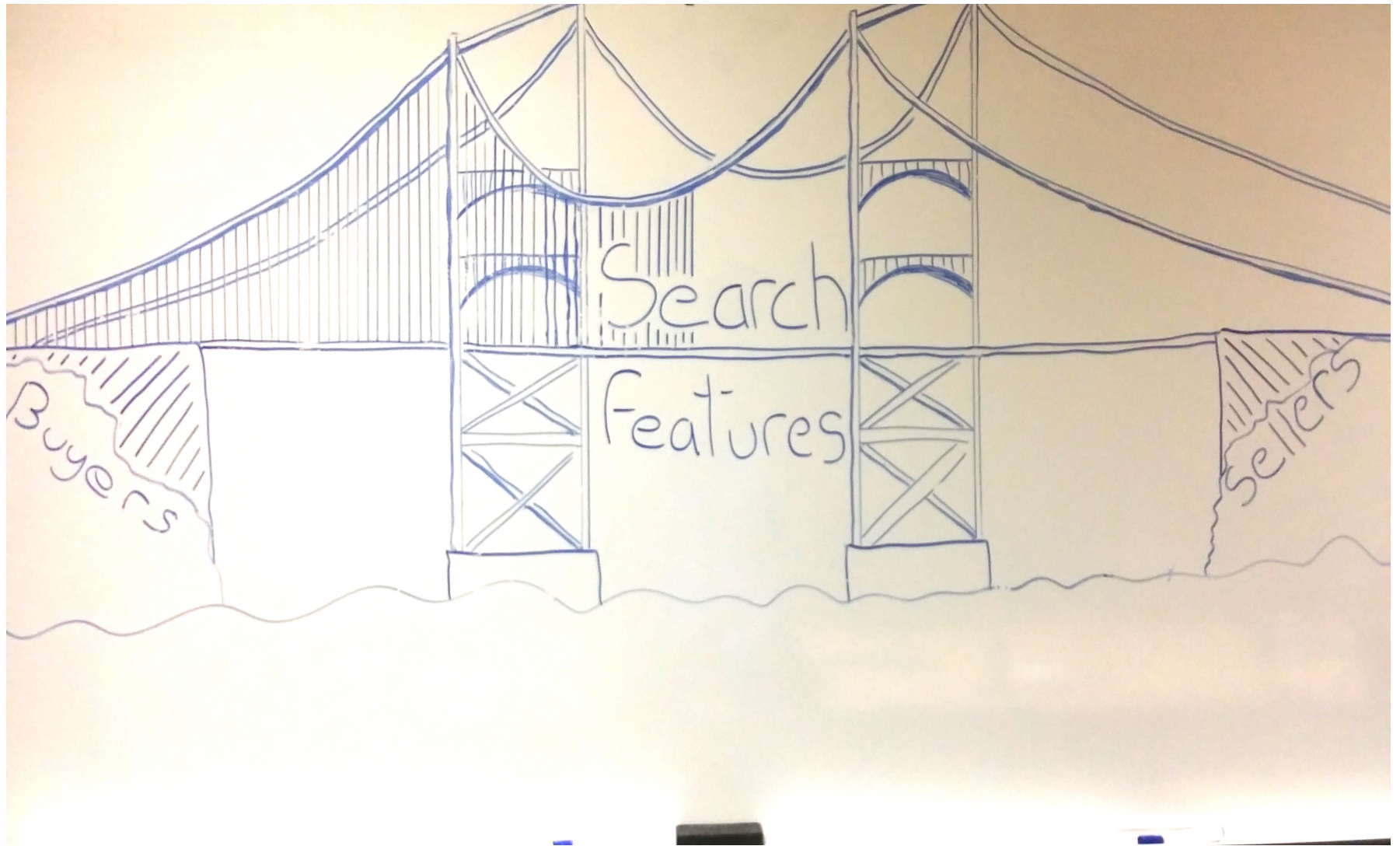


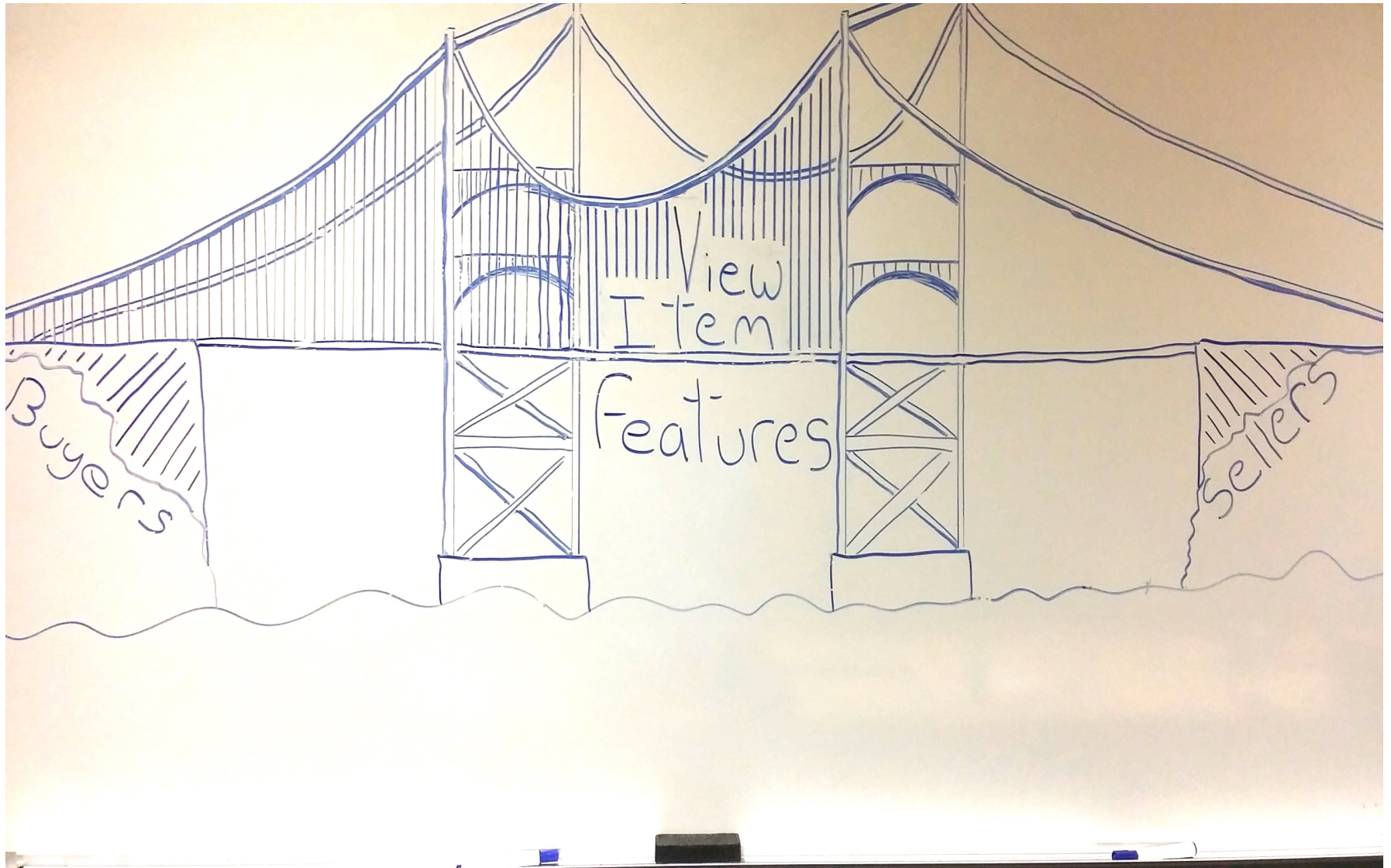


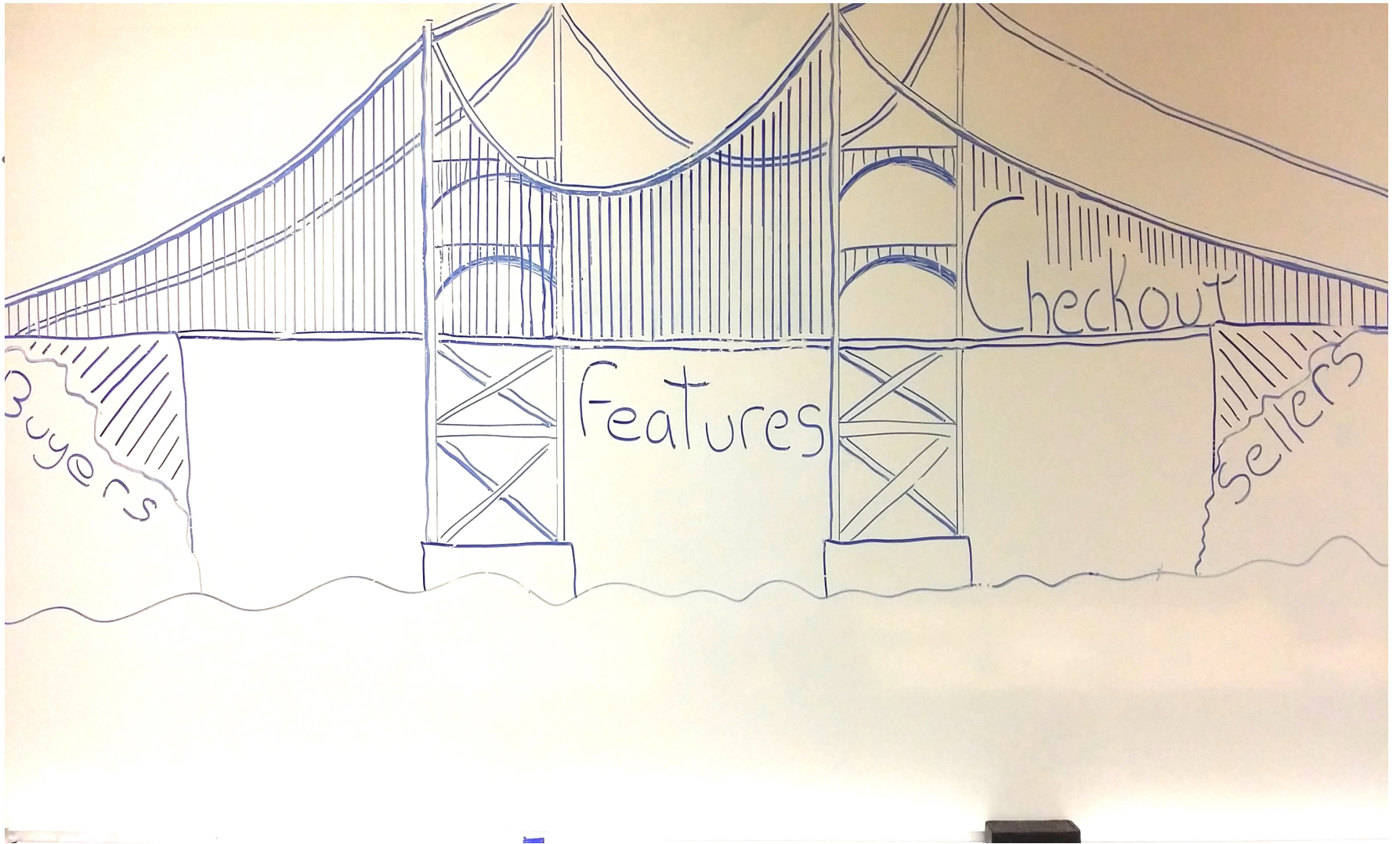


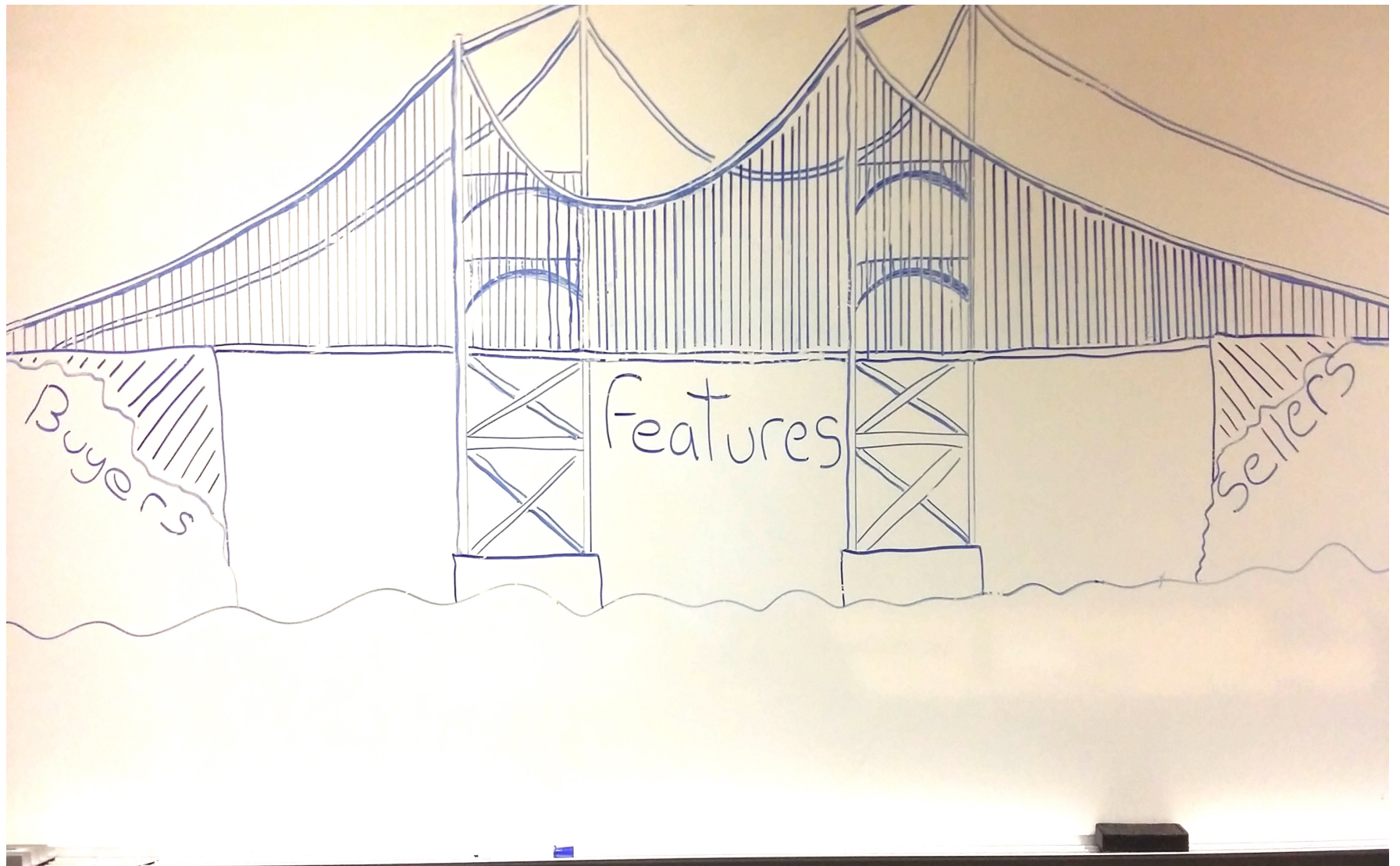


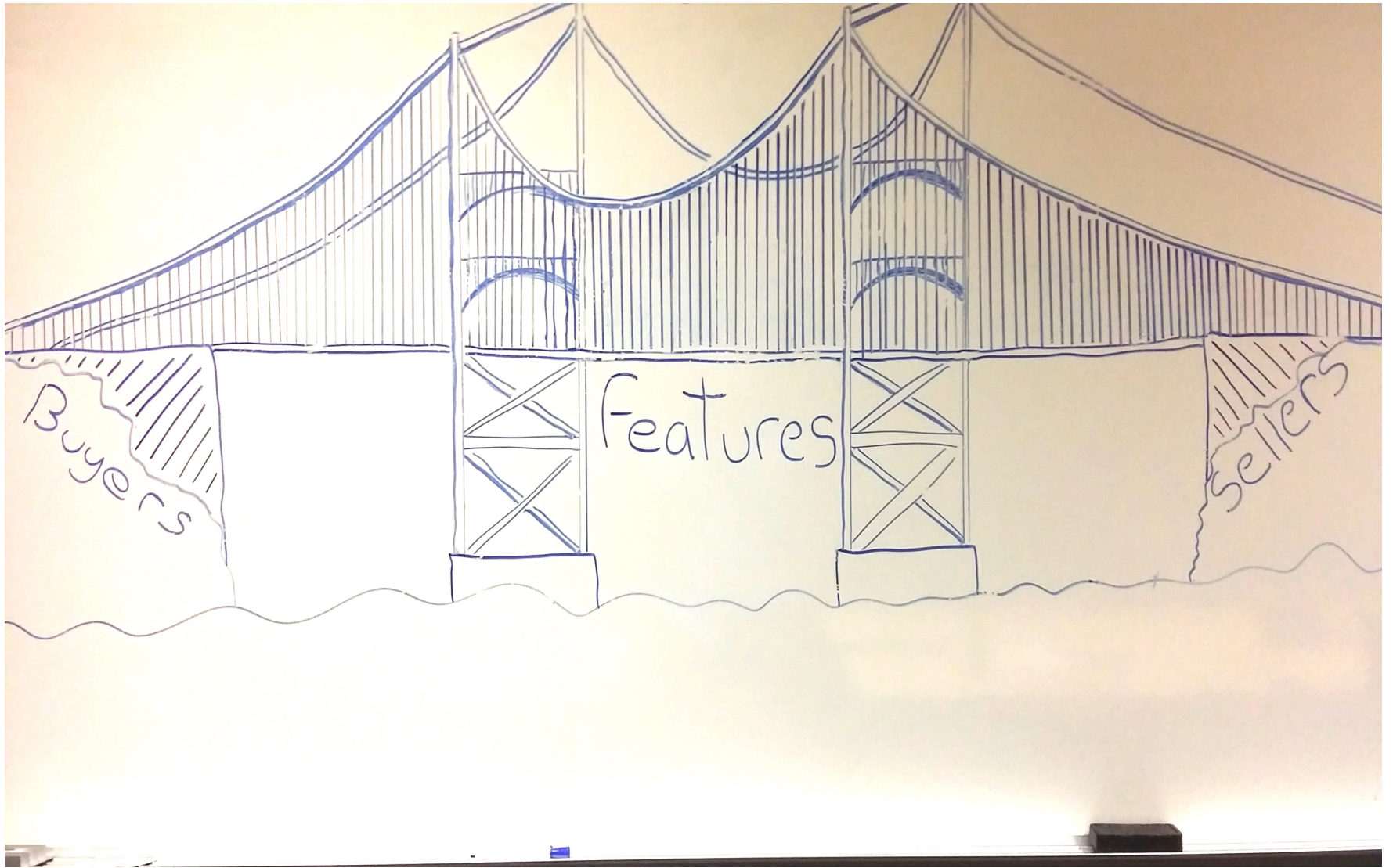


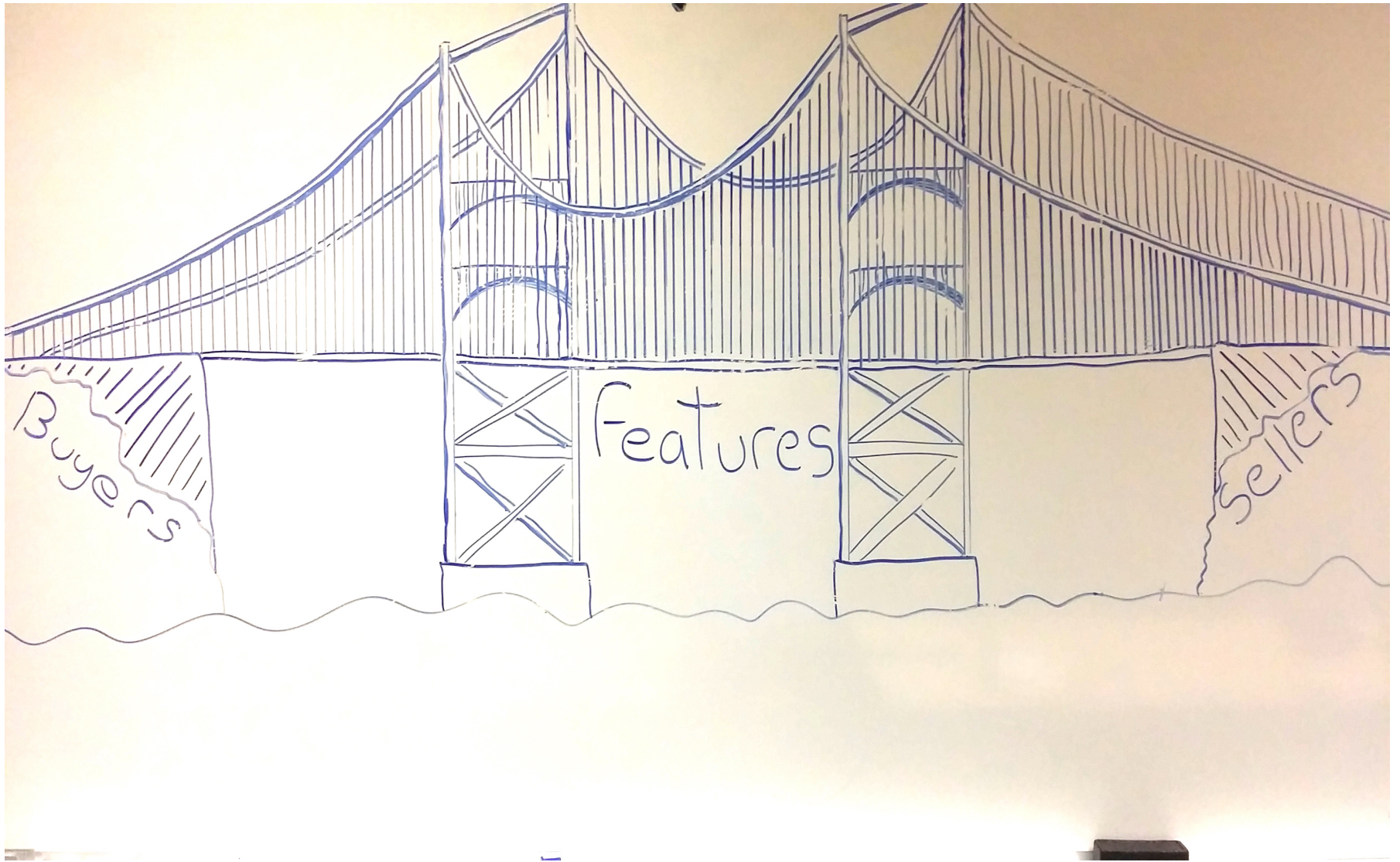




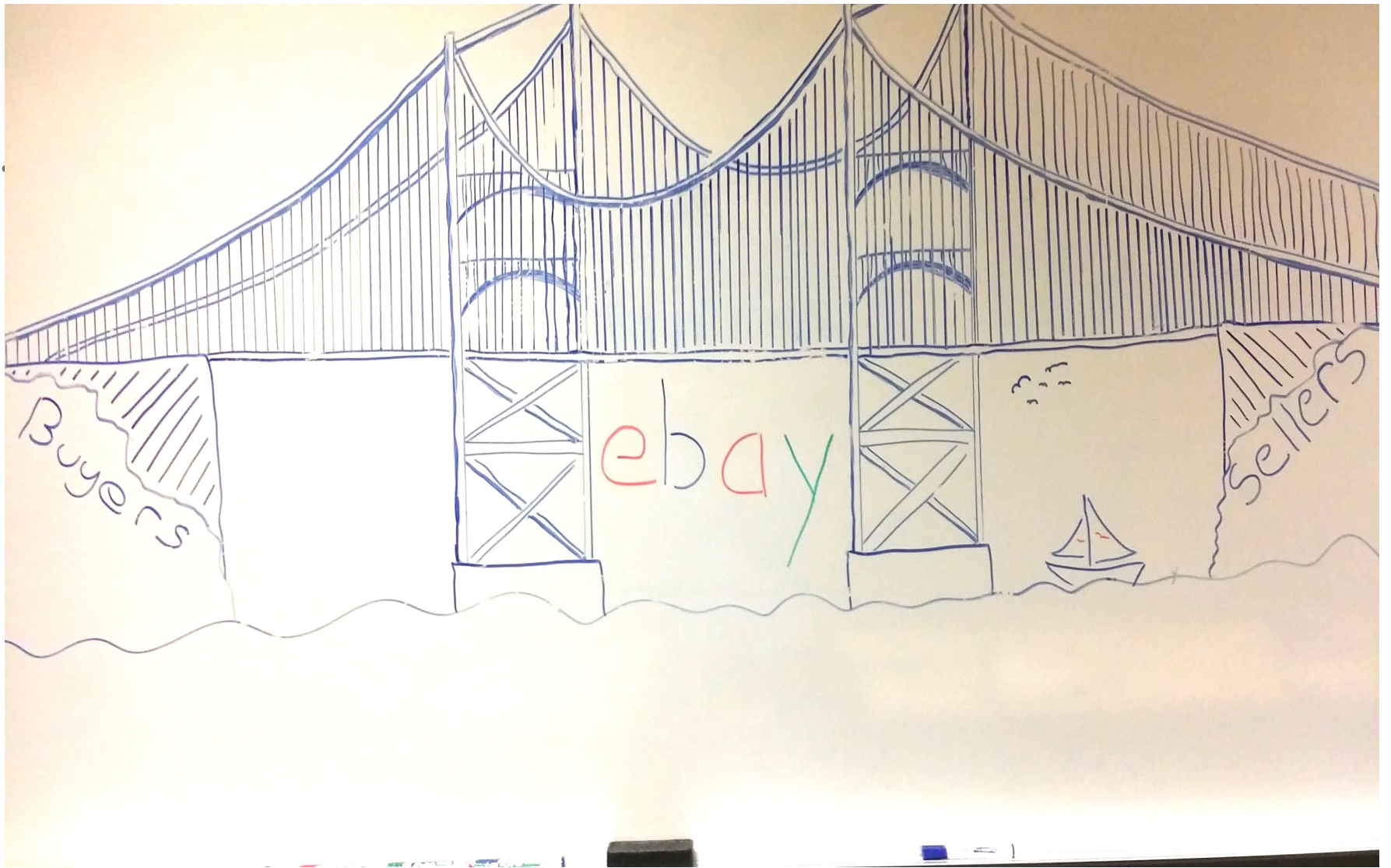












SO FULL OF FAIL

Creating resiliency through *intelligently* injected failure.

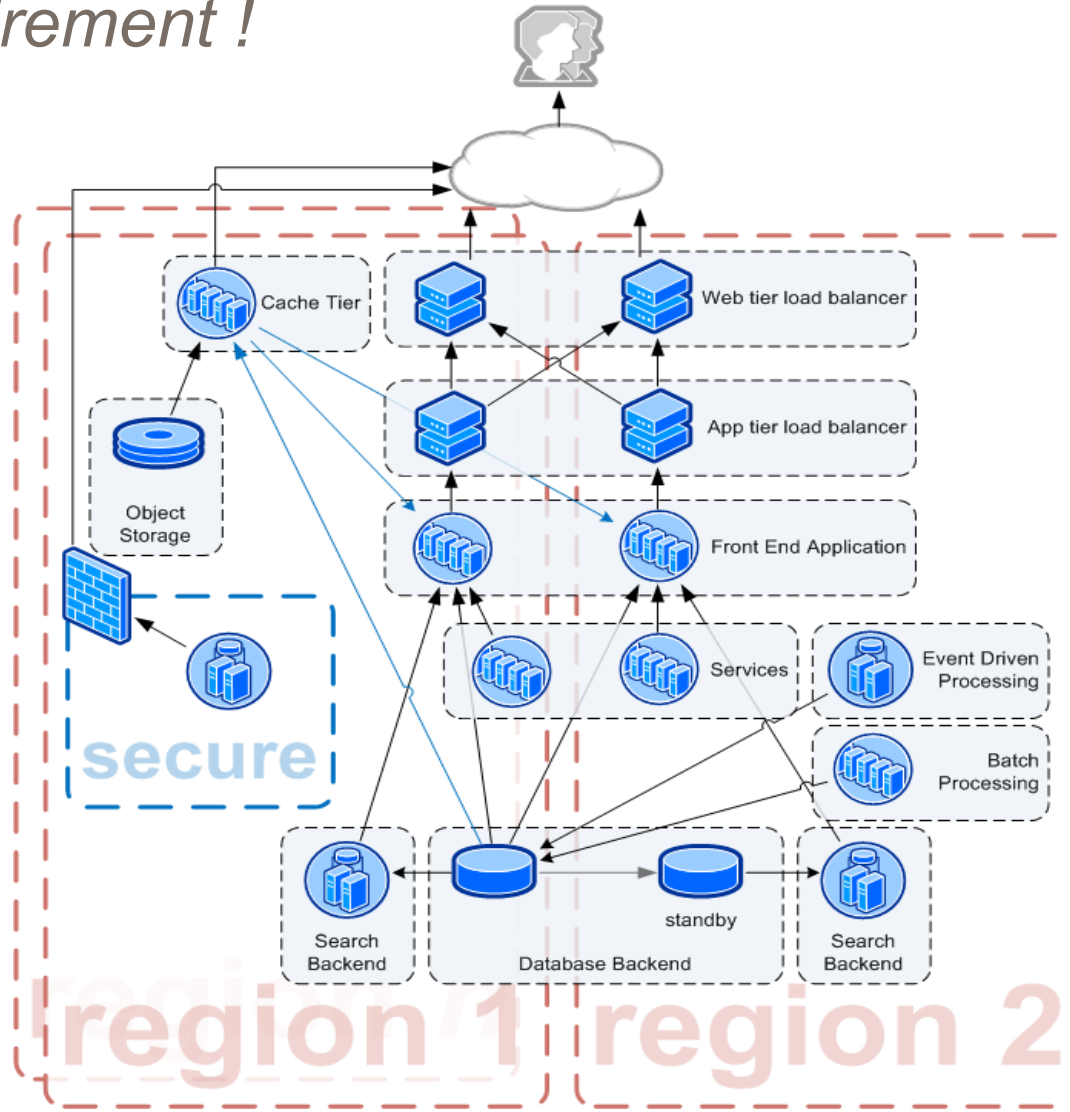


So Full of Fail – Failure is NOT an option. *It's a requirement !*

- The inability of a system or system component to perform a required function within specified limits. A failure may be produced when an error is encountered.



So Full of Fail – Failure is NOT an option.
It's a requirement !



So Full of Fail – Atomic failures – Overview

- **Software**

- Database failure
- Service failures

- **Hardware**

- LB failures
- Compute failures
- Network failures

So Full of Fail – Cascading failures

- A **cascading failure** is a failure in a system of interconnected parts in which the failure of a part can trigger the failure of successive parts or the whole system.

So Full of Fail – Cascading failures



So Full of Fail – Cascading failures



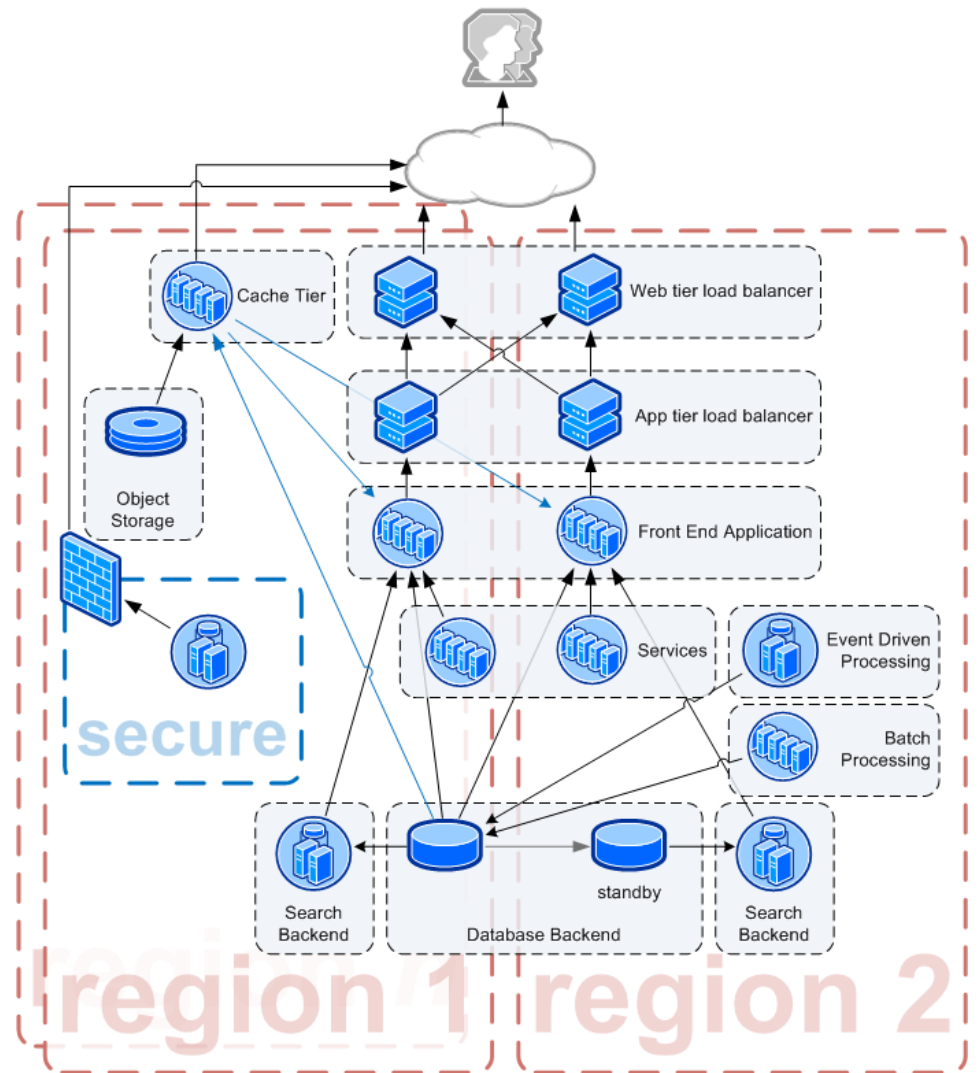
So Full of Fail – Cascading failures – Software

- **Service to Service**

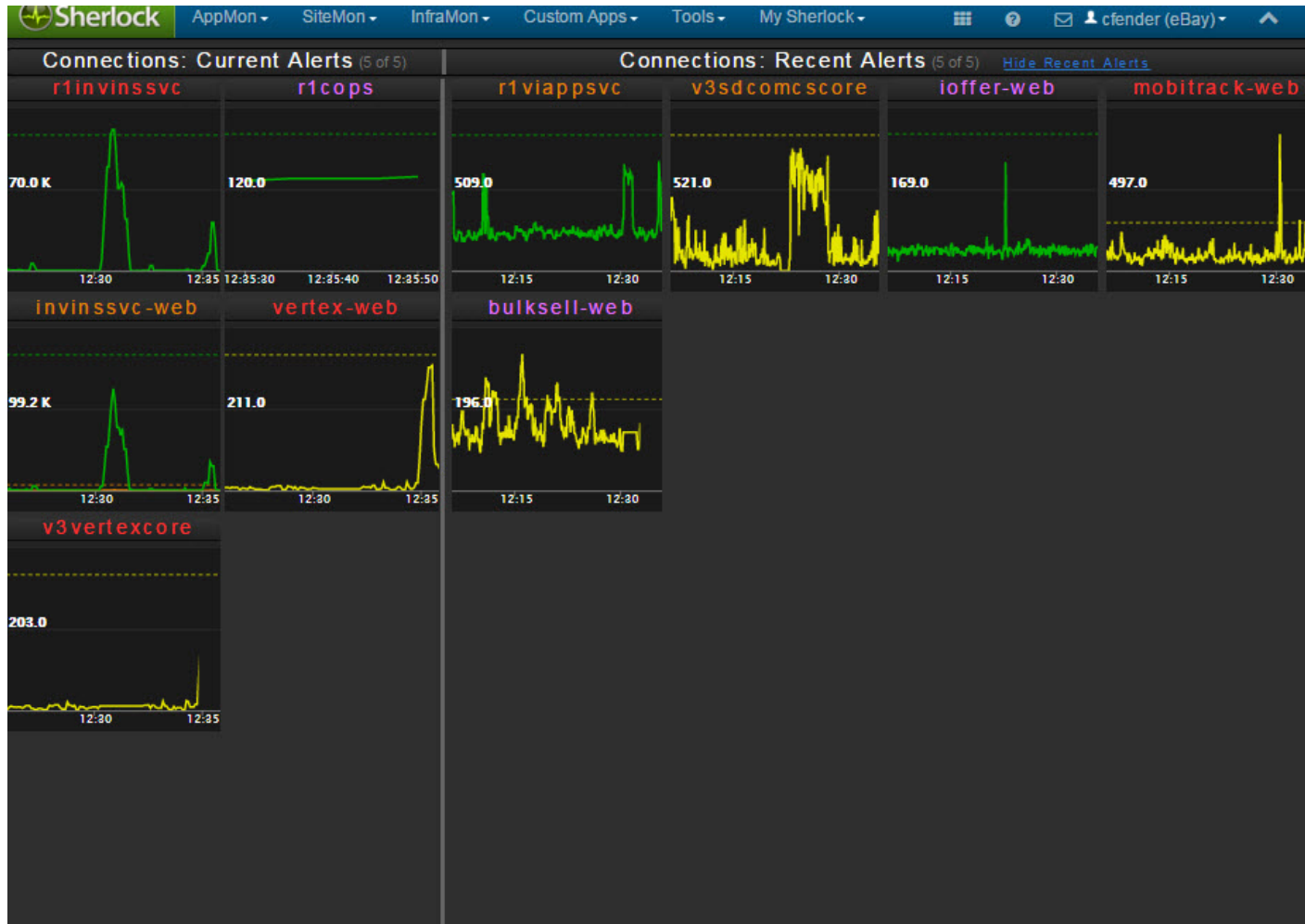
- Service A talks to Service B using using load balancer virtual ip

- **Service B to Database**

- Service B talks to data store using data access layer



So Full of Fail – Cascading failures – Software



So Full of Fail – Cascading failures – Software

Q Site Explorer Traffic: 8.409G | Code Roll: 60 (5) | Automation: 0 | User Actions: 1 (1) Last updated: Sat Jan 16 20:23:12. Next update in 53 (s)

Applications Alarms: 46/31 | Suppressed: 19192 | Snoozed: 0 Snooze Selected Restart Selected Nuke Selected Markup Selected Markup All +

SC MD: 41 | BES MD: 14 | APP HTTP: 28 | APP PING: 18 Class of Service Filter

half-agg-slc2-001	100% / 100%	H: 2	lhp-qry-phx-001	100% / 100%	H: 1	lhp-qry-phx-003	100% / 100%	H: 2
kemmisc-app	33% / 16%	H: 1	sunsetnj-app	33% / 12%	H: 1	emsvc-app	16% / 7%	P: 2
INFRA cmevpsqry-app	10% / 3%	H: 1	trackrmsi-app	9% / 5%	H: 1	dealcnsmr-app	4% / 4%	H: 2 P: 1
...iskConsumer.ORDER.FUNDING.UPDATE	BES MD: 5ngCart.TnsEvaluateService.HTTP11	SC MD: 4production.LPUpdateServiceClient	SC MD: 4	...
cif.sms.in.SMSClient2	SC MD: 3ioAsyncHttpClientConfig.HttpUtil	SC MD: 3le.services.util.TouchHttpClient	SC MD: 3	...
...fgsvc.production.GeoCfgSvcClient	SC MD: 3vc.production.lafAdmCosSvcClient	SC MD: 3viewUpdateConsumer.PRODUCT.MERGE	BES MD: 3	...

Databases Suppressed: 116 | Snoozed Markdowns: 4798 | Snoozed Databases: 3 Remediate Snooze Selected Markup Selected Markup All +

MD: 13 VCS Status: Faulted: 0 Partial: 2 Frozen: 4 Filter

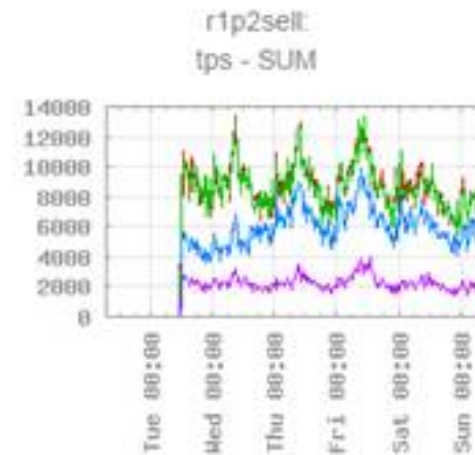
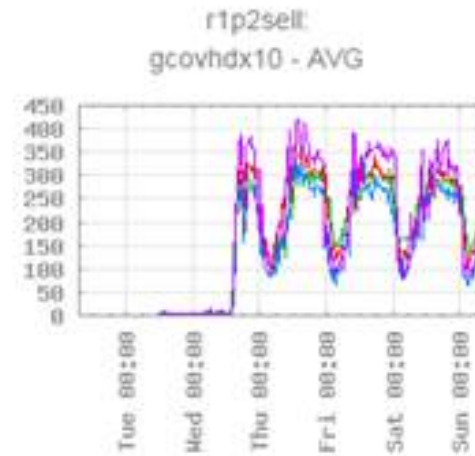
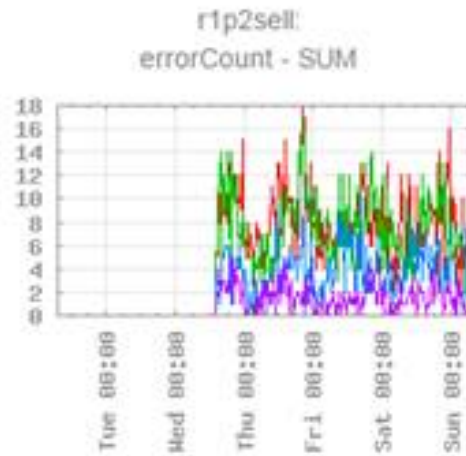
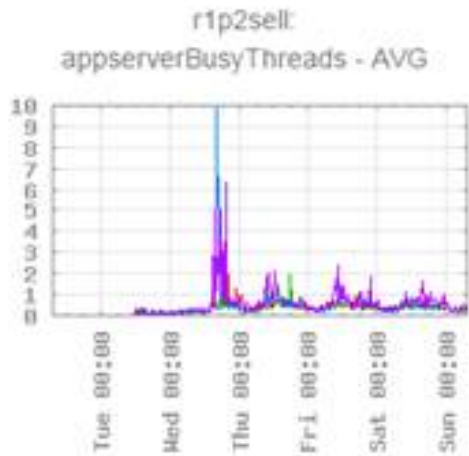
No alerts

T3 CORE_exadata01	MD: 9 +3	T1 ...okup__read-ulookup.sicrulookup04	MD: 1	T2 CORE_categoryhost18	MD: 1
T3 CORE_bes0	MD: 1	T3 CORE_myebaylookuphost	MD: 1		

Actions +

Configuration data powered by CMS. Montage UI Version: 15.0

So Full of Fail – Cascading failures– Software



So Full of Fail – Cascading failures – Software - Database

- Connectivity failure
- Query timeouts
- SQL errors

The screenshot displays the MONTAGE monitoring interface. At the top, it shows system metrics: Traffic: 14.900G, Code Roll: 63 (11), Automation: 9 (5), and User Actions: 1425 (21). Below this, it indicates 'Databases Suppressed: 492 | Snoozed Markdowns: 3697 | Snoozed Databases: 5'. A status bar shows 'VCS Status: Faulted: 0 Partial: 0 Frozen: 2'. Two database entries are highlighted in yellow:

T3	All	7m	Mongo-lbmslv...	lvs2b02c-ab...	8.15	63%	5	T3	All	6h	cass-duvs-phx04	caduvsphx04..19.81	174	9%
----	-----	----	-----------------	----------------	------	-----	---	----	-----	----	-----------------	--------------------	-----	----

At the bottom of the dashboard, a red message states 'No markdown alerts'.

So Full of Fail – Cascading failures – Software

- Data access layer identifies any of the above issues and marks the connection down.
- When the destination is recovered the path is marked active.
- Time to detect the path state is ~15 ms
- Request for connections to a struggling DB can lead to to stacking in upstream
- The fail fast waiter queue logic helps the loaded database

So Full of Fail – Cascading failures – Software



Administration Console

Host: lvsviewitem-86169/10.134.108.134 (client IP 10.28.111.89)

- Configuration
- Component Status
- Service Client Status
- JMX MBeans
- Hystrix Dashboard
- Logs

ebay.kernel.ServeTraffic		Filter:
Configuration	Value	
Alias	ebay.kernel.ecv	
Description	Serve Traffic to this server.	
Group	ebay.kernel	
ID	ebay.kernel.ServeTraffic	
Initializable		
Last Updated	Thu Mar 24 14:38:04 GMT-07:00 2016	
Persistent	true	
Persistent Location	file:/ebay/cronus/software/service_nodes/.ENV85r9jj2wb0.viewitem-app__ENV85r9jj2wb0.viewitem-app__ENV85r9jj2wb0-LVS-CLhyz5f6oi39d2kg-10.134.108.134/installed-packages/Tomcat/7.0.47_13_raptor_taginstance.unx/cronus/scripts/Tomcat/webapps/ROOT/WEB-INF/classes/appconfig/Production/raptorconfig/config/temp_persist_config_015.xml	
Site Operations Command	true	
Validator Class Name	com.ebay.kernel.bean.configuration.adapter.ConfigManagementValidatorAdapter	

Click the below values to edit

- ▶ Current changes
- ✖ Value not updated
- ✔ Value updated

Configurable Properties		Filter:
Property	Value	
Value	TrafficEnabled	

Add property Reset Property Submit Cancel



So Full of Fail – Cascading failures – Software

Command & Control ✕

[Database Markup](#) **Database Markdown** [Database SwitchOver](#) [Database FailOver](#) [Continuous Markup](#)

Database Name	Application Service Selection	Application Services	Node Servers	Sitewide (Ignore DBMap)
caty10stby	<input type="text" value="all application services"/>	<input type="text" value="Search..."/> <div style="border: 1px solid #ccc; height: 40px; width: 100%;"></div>	<div style="border: 1px solid #ccc; background-color: #f0f0f0; height: 100%; width: 100%;"></div>	<input type="text" value="No"/> ✕

Submit



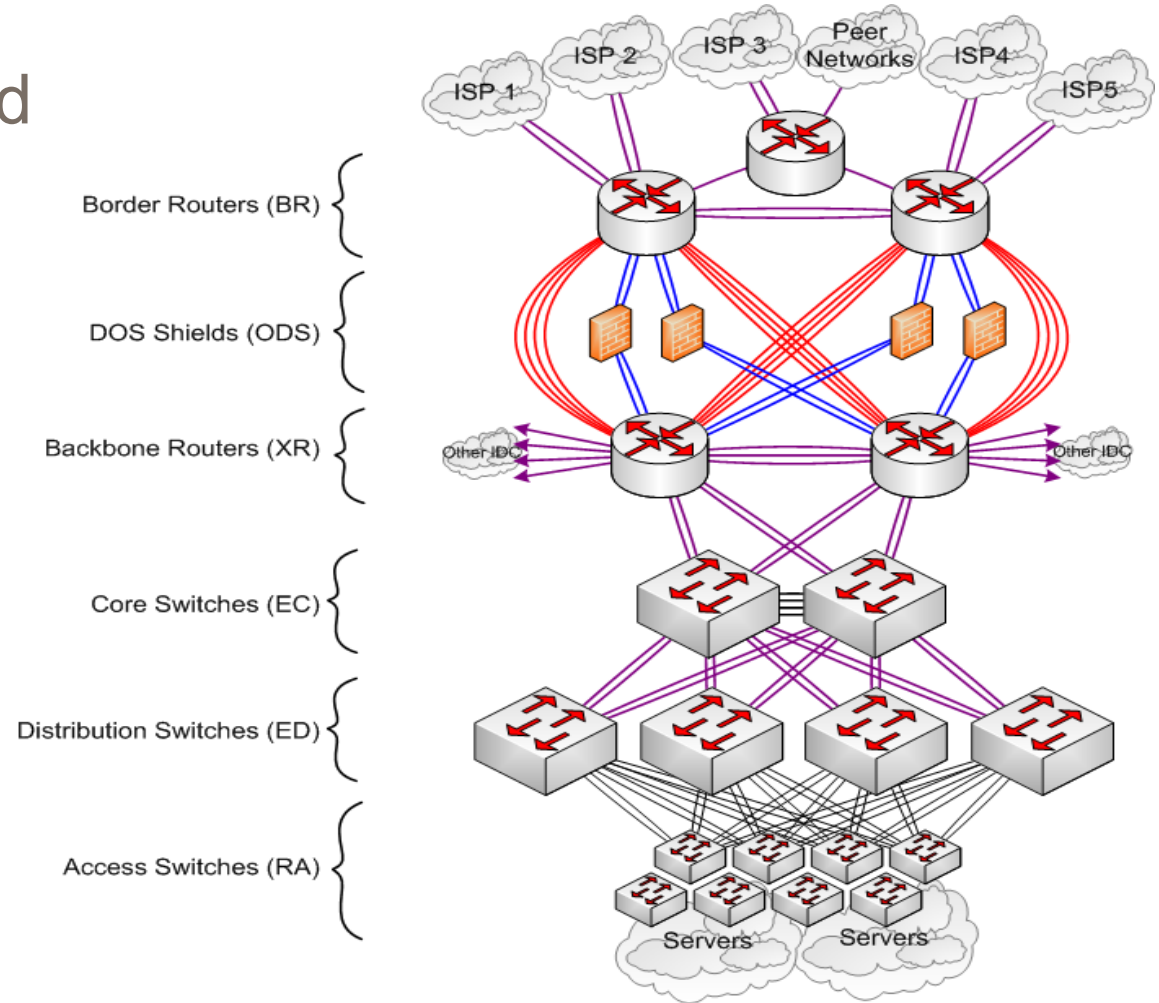
So Full of Fail – Cascading failures – Hardware

- LB failures
- Compute failures
- Network failures

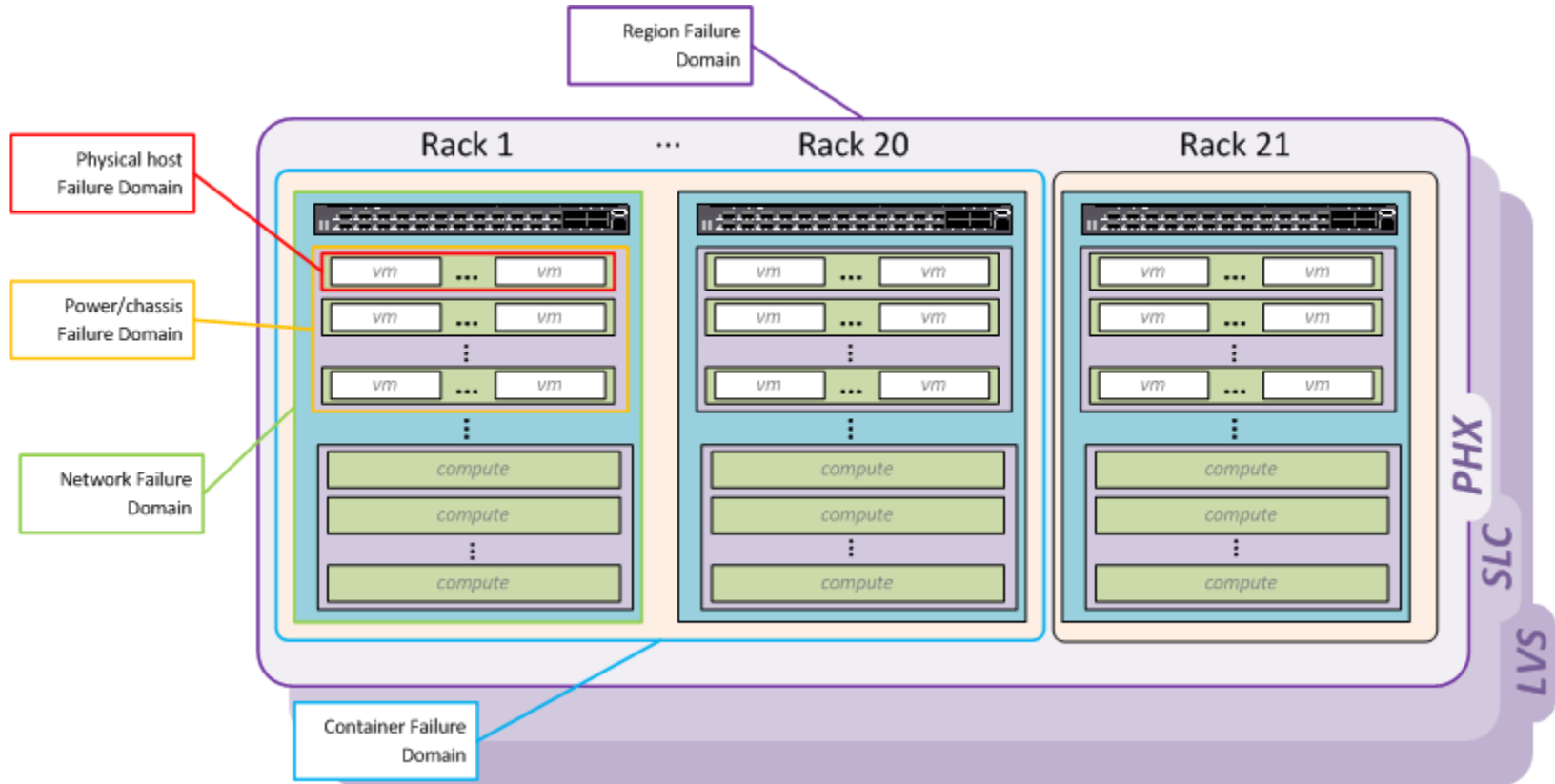


So Full of Fail – Cascading failures – Hardware - Network

- Traffic organized in layers.
- Redundant interconnected paths.



So Full of Fail – Cascading failures – Hardware



$$P(n) = \text{MAX} \left(17\%, \frac{1}{S_n}, \frac{1}{F_n} \right)$$

- n = Failure Domain (e.g. Network)
- S_n = Number of Service Instances in Failure Domain n
- F_n = Number of instances of Failure Domain n

PREVENTION THROUGH PARADIGM

Prevent and remediate atomic failure in order to prevent cascading failure scenarios.

PREVENTION THROUGH PARADIGM

- Site code is deployed on a common set of platforms which enforce the aforementioned tools.
 - Engineering and development architecture effort is “front-loaded” to ensure that all production services can be restarted in exactly the same way no matter what the underlying command.
 - Uniformity of response
 - Remediation at the node level can be automated.
- State based and event based monitoring of key operating metrics.
 - The 9 or so OS-level metrics we all monitor
- Capacity monitoring for “DR’ compliance
 - Two or more co-locations
 - One feature per “pool”
- Provision on demand
- Regional code roll

PARADIGM BEFORE PROCEDURES.

- For maximum rapid remediation your pool or feature should comply with operational paradigms.
- If the care and feeding of your particular beautiful 'unicorn' requires special attention or procedures:
 - You're asking someone to go against their established paradigm and training.
 - At a high rate of speed
 - As the rarely seen or remembered exception



DEALING WITH DISASTER DIRECTION AND DECISION

Preventing the cascade failure through redirection and deciding which features to keep.

When a lion tamer holds a chair in front of the lion's face, the lion tries to focus on all four legs of the chair at the same time. With its focus divided, the lion becomes confused and is unsure about what to do next. When faced with so many options, the lion chooses to freeze and wait instead of attacking the man holding the chair.



Clyde Beatty taming a lion with a chair. (Image from [Harvard Library](#).)

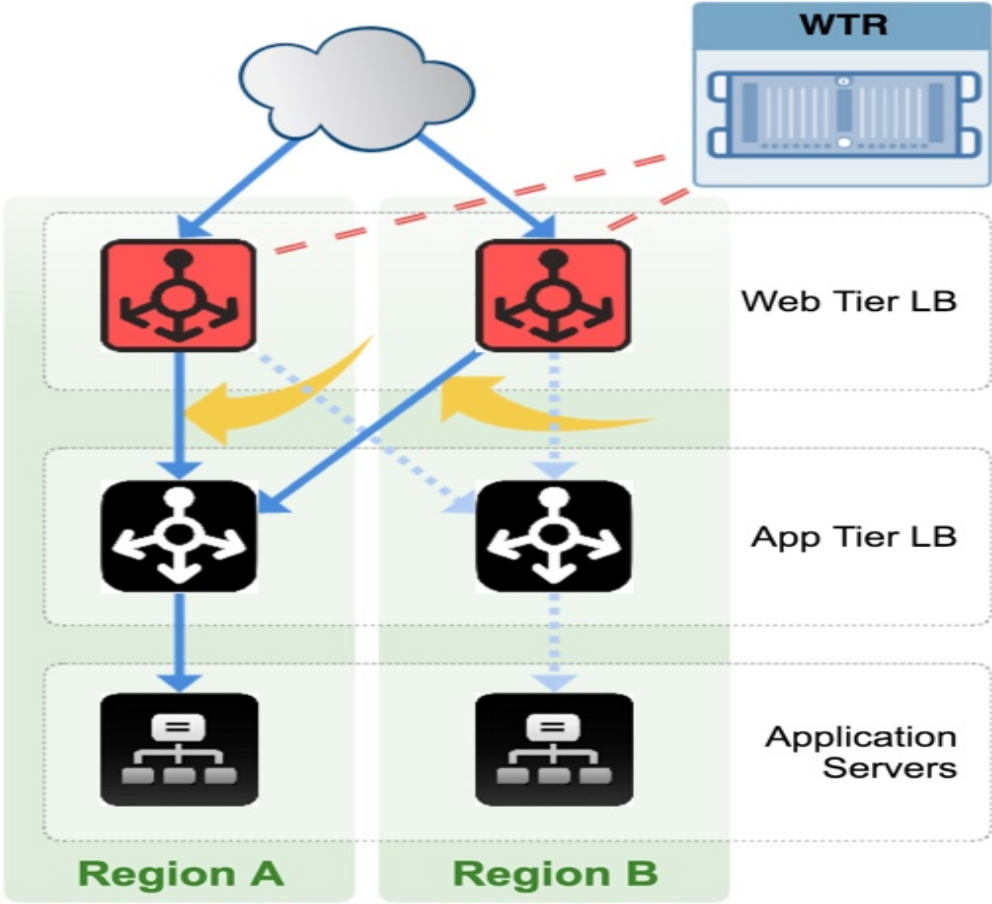








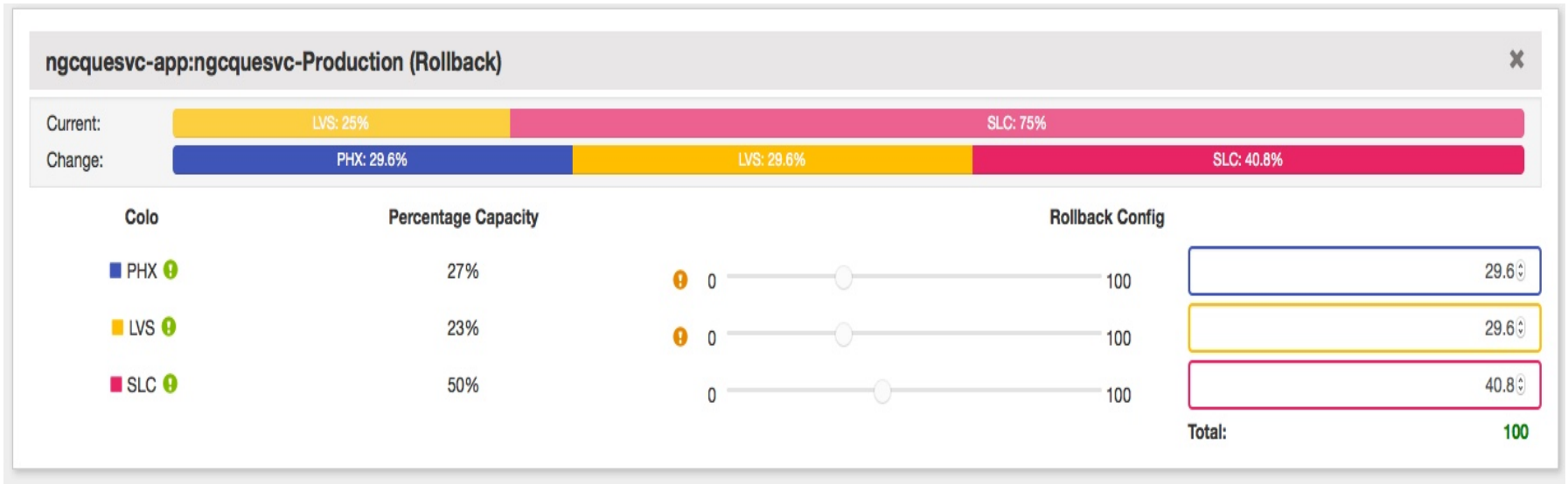
DEALING WITH DISASTER - DIRECTION



DEALING WITH DISASTER - DIRECTION



DEALING WITH DISASTER - DIRECTION



DEALING WITH DISASTER - DIRECTION

The screenshot displays a disaster recovery management interface with three panels, each representing a different LVS (Load Balancing Service) configuration. Each panel includes a title, a description of the service, a status indicator, and a list of items with checkboxes and buttons for bulk actions.

Titan LVS

Serving SLC
Status: In partial DR

Item	Checkbox	Action	Source
activeitem_slc_qn	<input type="checkbox"/>	bulk	From LVS
	<input type="checkbox"/>	mini	From LVS
activeitem3_slc_qn	<input type="checkbox"/>	bulk	From PHX
	<input type="checkbox"/>	mini	From LVS
completed_slc_qn	<input type="checkbox"/>	bulk	From LVS
	<input type="checkbox"/>	mini	From LVS

Titan PHX

Serving CHD and PHX
Status: Normal

Item	Checkbox	Action	Source
activeitem_chd_qn	<input checked="" type="checkbox"/>	bulk	From PHX
	<input checked="" type="checkbox"/>	mini	From PHX
activeitem3_chd_qn	<input checked="" type="checkbox"/>	bulk	From PHX
	<input checked="" type="checkbox"/>	mini	From PHX
activeitem3_phx_qn	<input checked="" type="checkbox"/>	bulk	From PHX
	<input checked="" type="checkbox"/>	mini	From PHX

Huygens LVS

Serving LVS
Status: Normal

Item	Checkbox	Action	Source
activeitem_lvs_qn	<input checked="" type="checkbox"/>	bulk	From HUYGENSLVS
	<input checked="" type="checkbox"/>	mini	From HUYGENSLVS
activeitem3_lvs_qn	<input checked="" type="checkbox"/>	bulk	From HUYGENSLVS
	<input checked="" type="checkbox"/>	mini	From HUYGENSLVS
completed_lvs_qn	<input checked="" type="checkbox"/>	bulk	From HUYGENSLVS
	<input checked="" type="checkbox"/>	mini	From HUYGENSLVS

DEALING WITH DISASTER - DIRECTION

- To prevent total failure of the bridge we have to get the kittens to drive over only one lane
- Connection limits would be required
- Blocking bots
- Scale down to minimal serving modes
- Wiring off features like advertising.



CONCLUSIONS

So Full of Fail– Total Failure Avoidance

- Bubble up anomalies for inspection
- Fail Fast
- Mark down the failed paths
- Preserve the user experiences
- Use system approaches to solve