Reliable RT processing @ Spotify

Pablo Barrera <pablo@spotify.com>

Spotify®

February 5, 2014

Spotify

Spotify

- the right music for every moment
- over **6 million paying customers**
- over **24 million active users** each month
- over **20 million songs**
- over **1.5 billion playlists** created so far
- available in **55 markets**

i/o tribe

responsible for building the awesome infrastructure that supports the Spotify experience

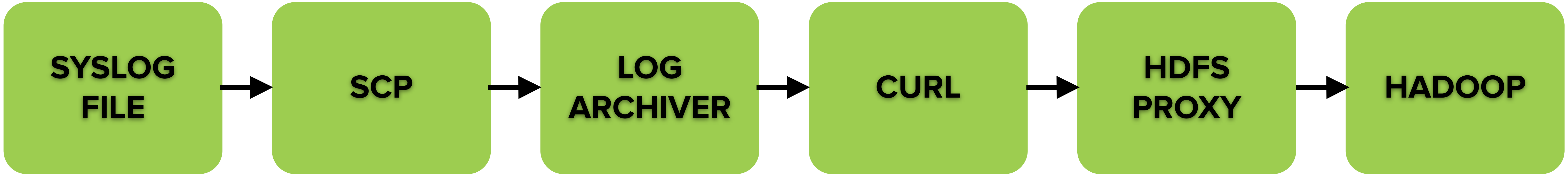
Our goal

this looks easy

easy
STAPLES®

but we have a problem...

Naïve approach (tm)



SYSLOG FILE → SCP → LOG ARCHIVER → CURL → HDFS PROXY → HADOOP

SCP

CURL

SCP

CURL

Scalability

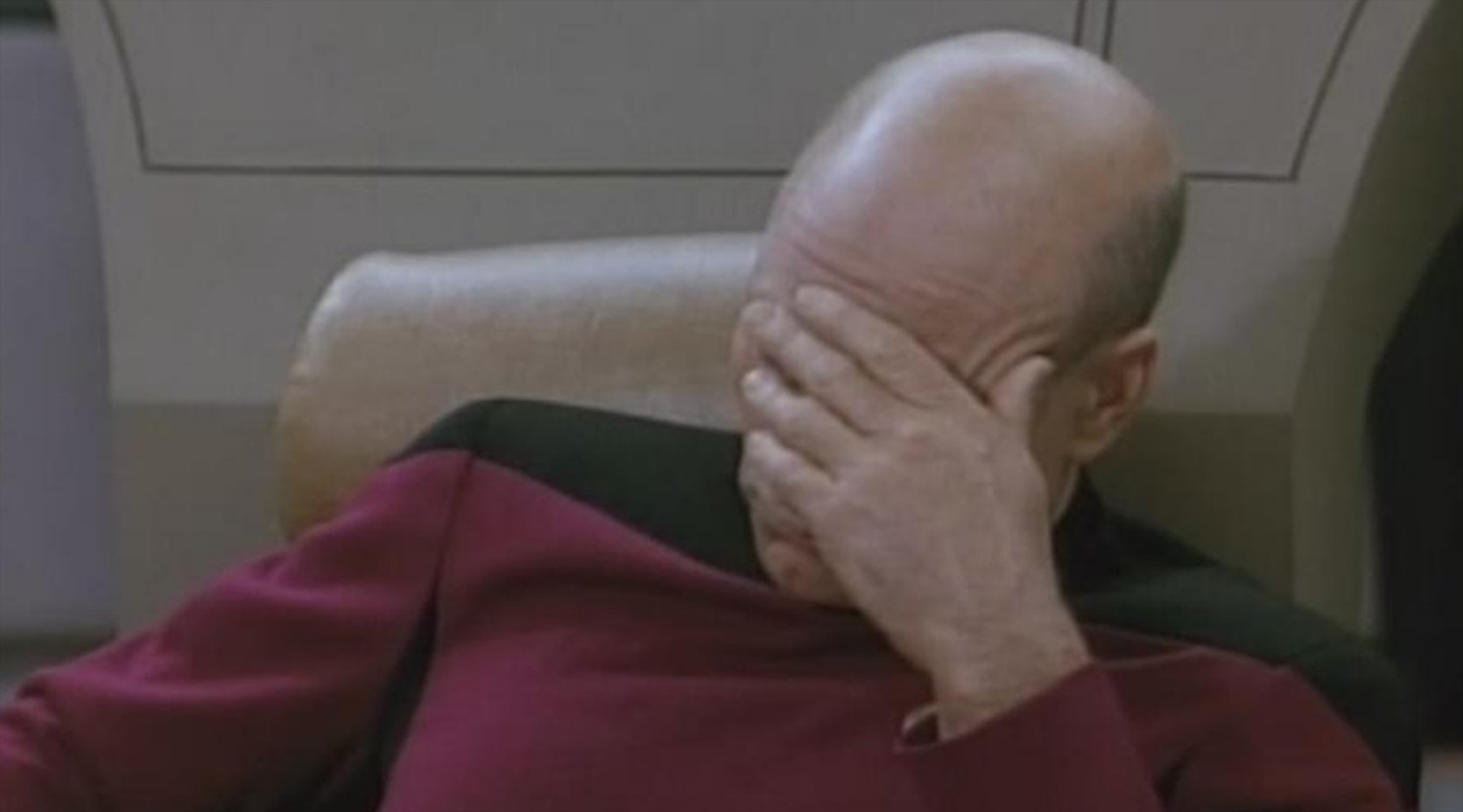
SCP

CURL

Scalability

```
for(;;) {
    SCP (file)
}
```

```
for(;;) {
    CURL (file)
}
```



thousands of servers
several data centres
millions of users

10 TB

each day!

Our Needs

- reliable delivery
- fast data transfer
- per-service subscription
- low cpu overhead

EC5274
SS/SD
ИЗОТ
СДЕЛАНО В БОЛГАРИИ
boedem
FLOPPY DISK
5,25'' HD,DS
BASF
HD
WRITE ENABLE
WRITE PROTECT
INDEX
HD

Other options

- active mq/rabbit mq
- flume/flume-ng
- others: scribe, chukwa, bookkeeper

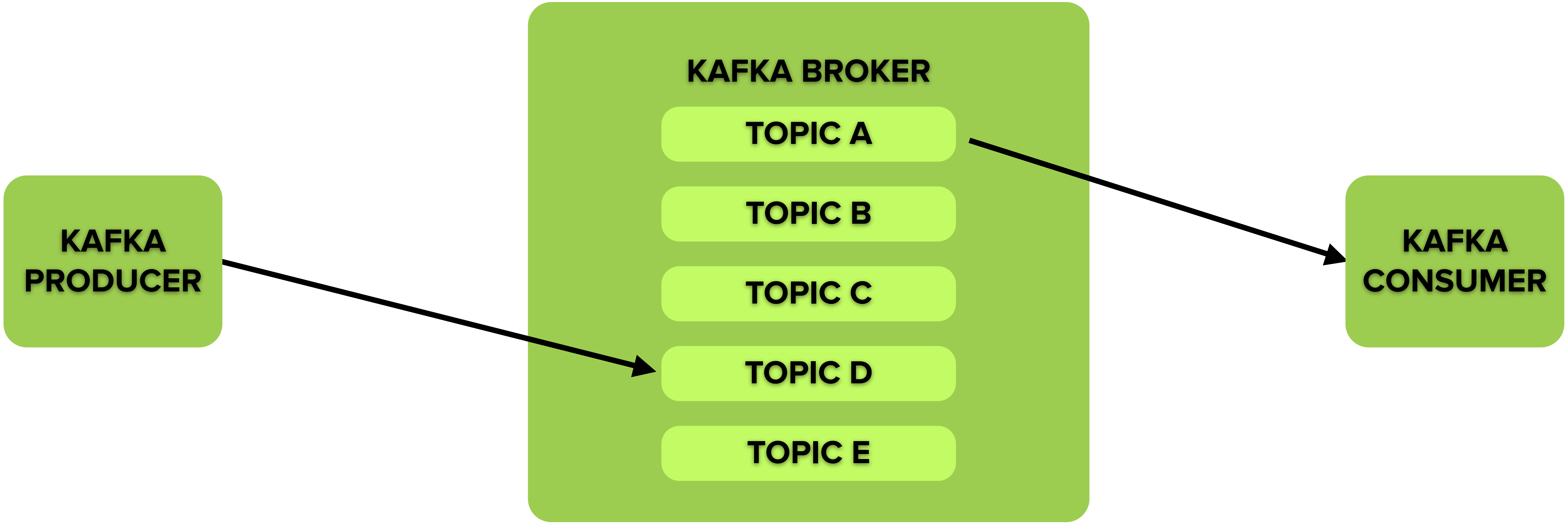
Apache Kafka

distributed pub/sub system

Kafka coolness

- at least once read
- O(1)
- network bounded

Kafka architecture



Cons

- no reliability
- no replication
- manual tuning

Spotify <3 Kafka

running in production!

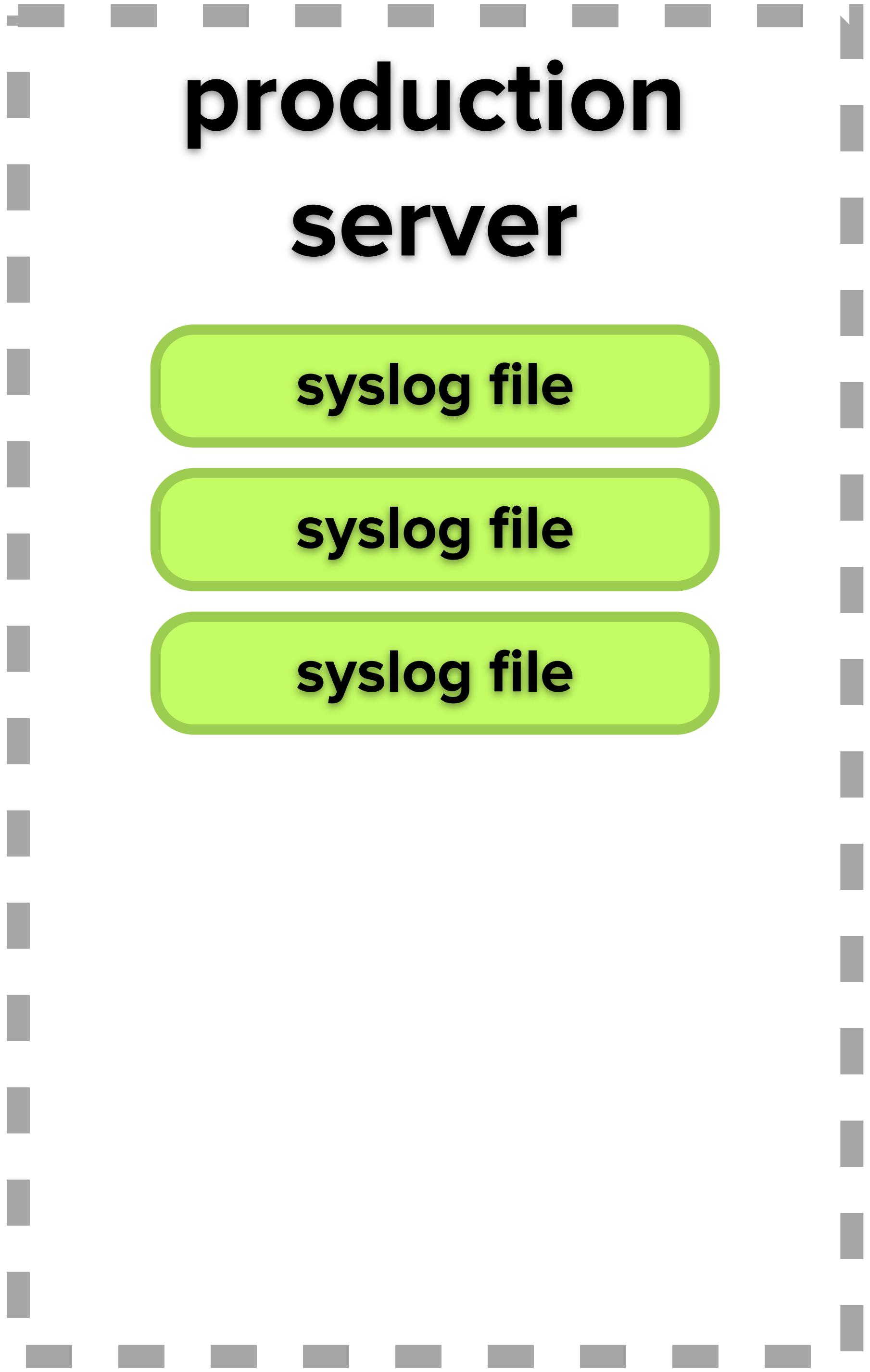
Kafka at Spotify

- key component of our log delivery system
- kafka 0.7.1
- java 7

Custom extensions

- end-to-end reliable delivery
- compression/encryption service

End-to-end reliable delivery

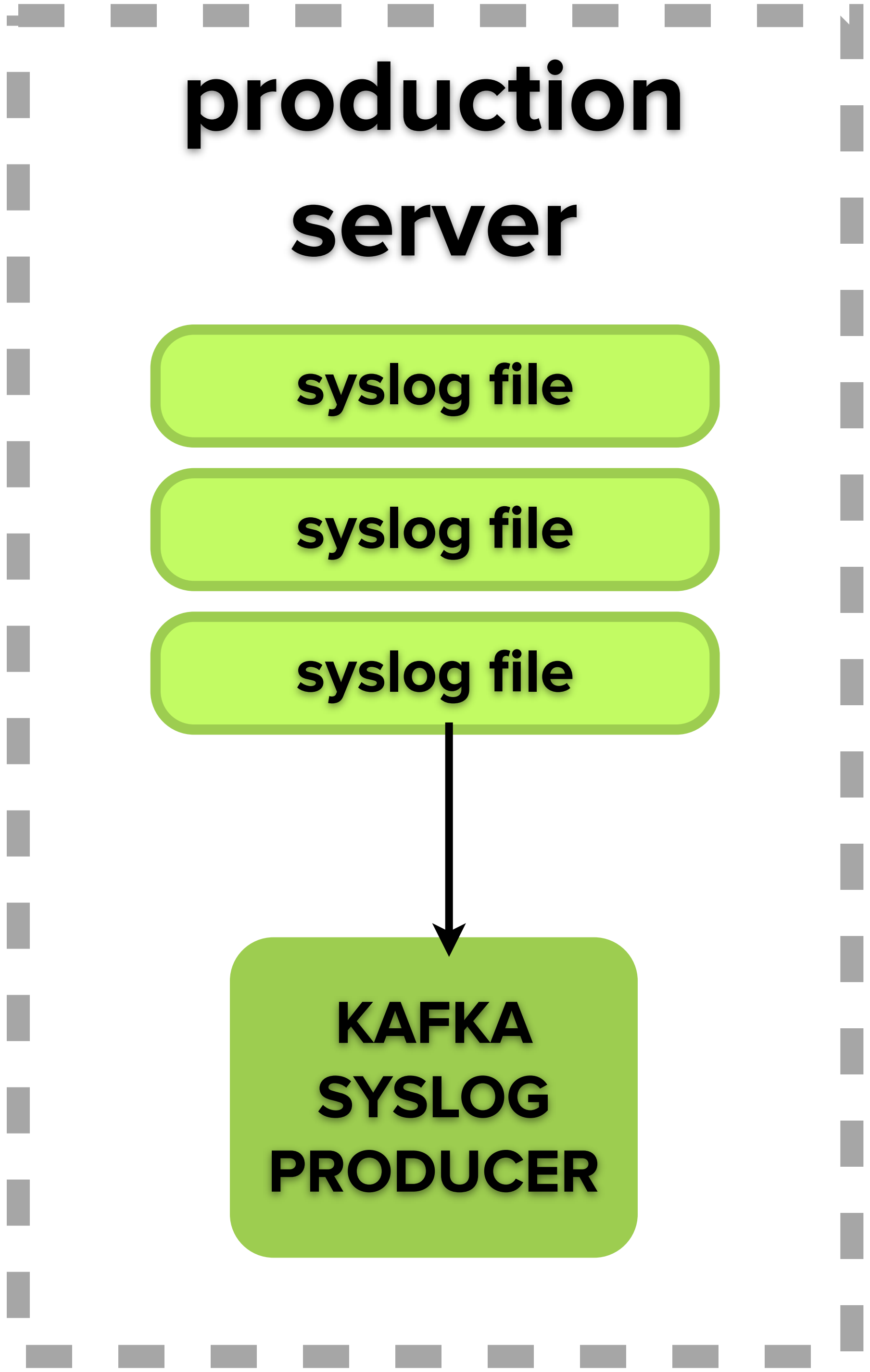


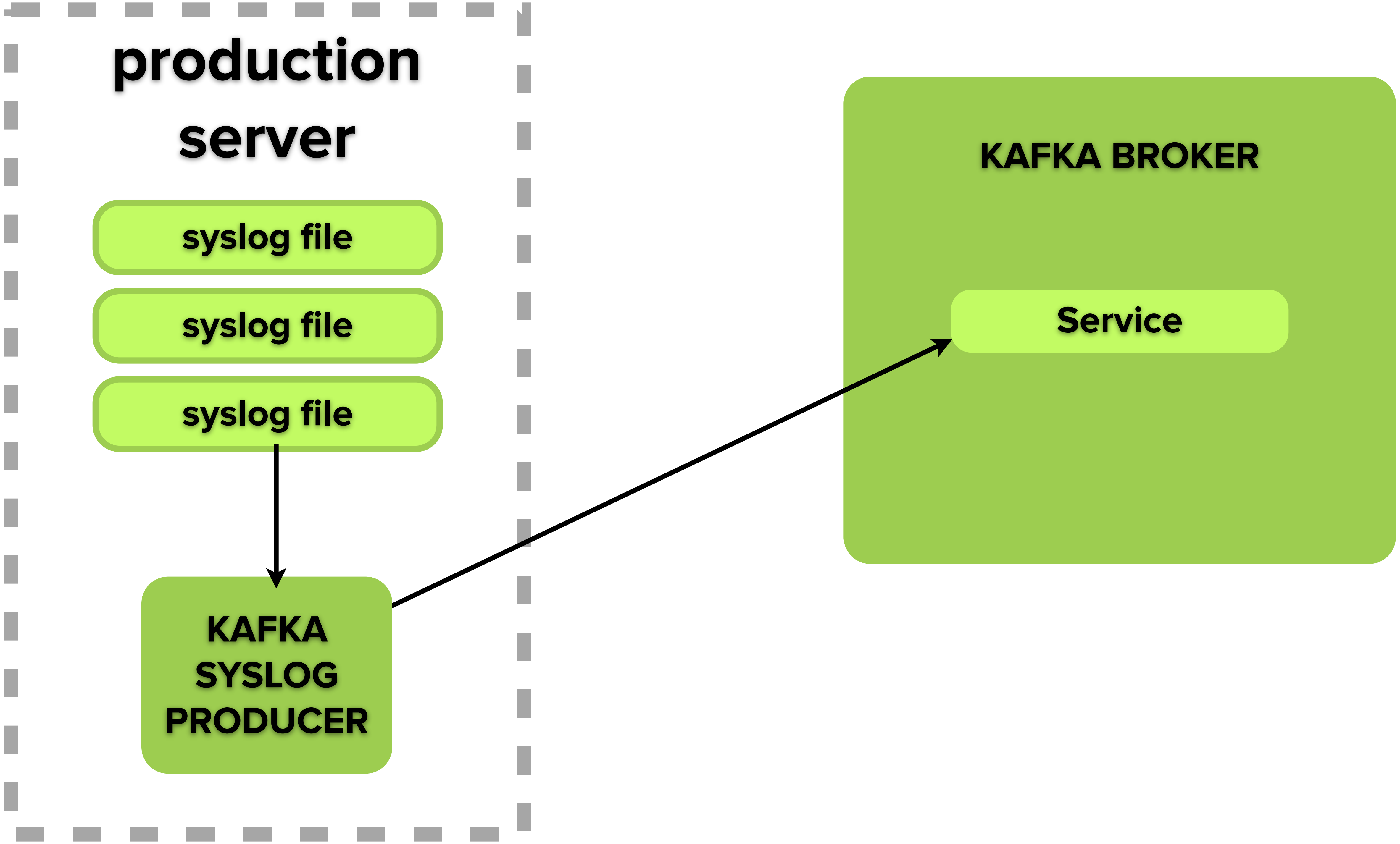
production server

syslog file

syslog file

syslog file


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER



production server

syslog file

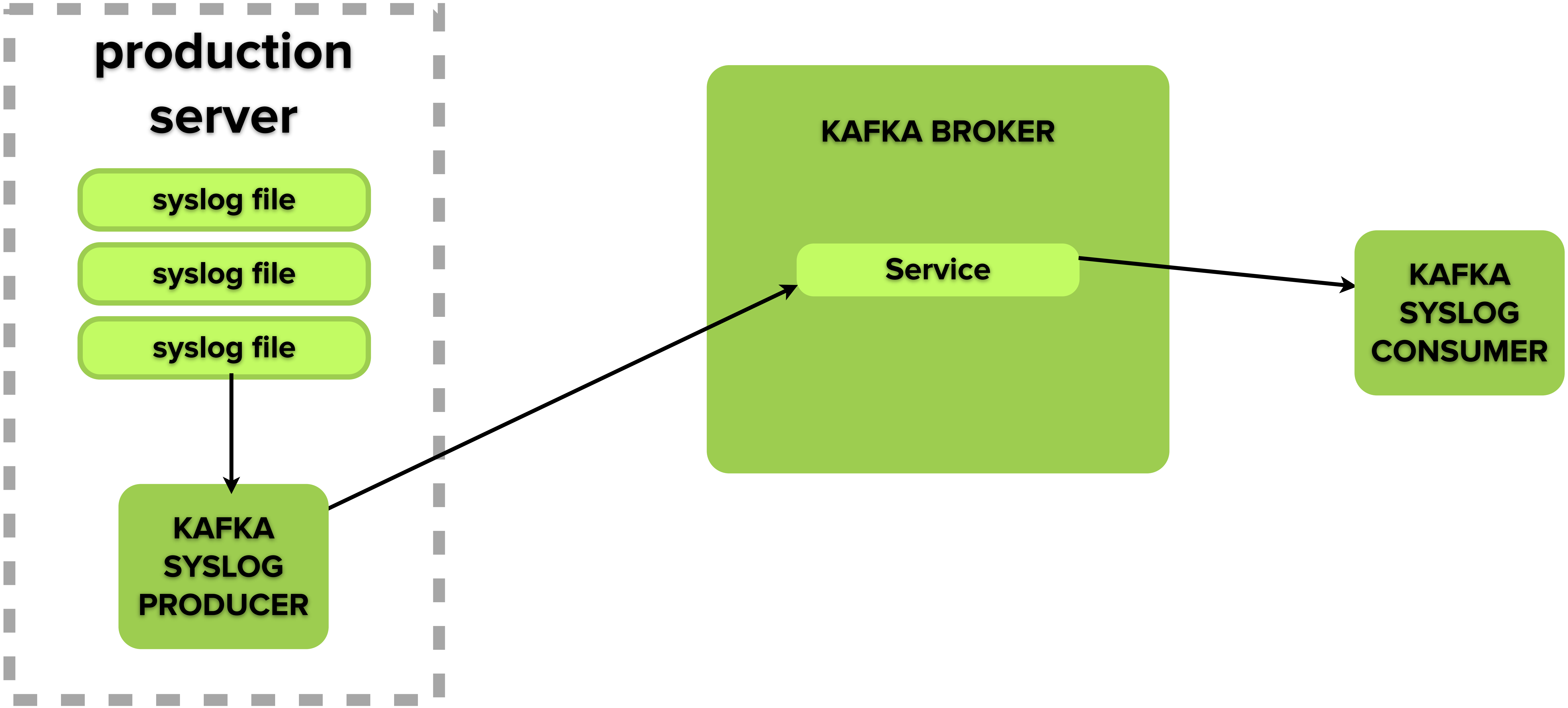
syslog file

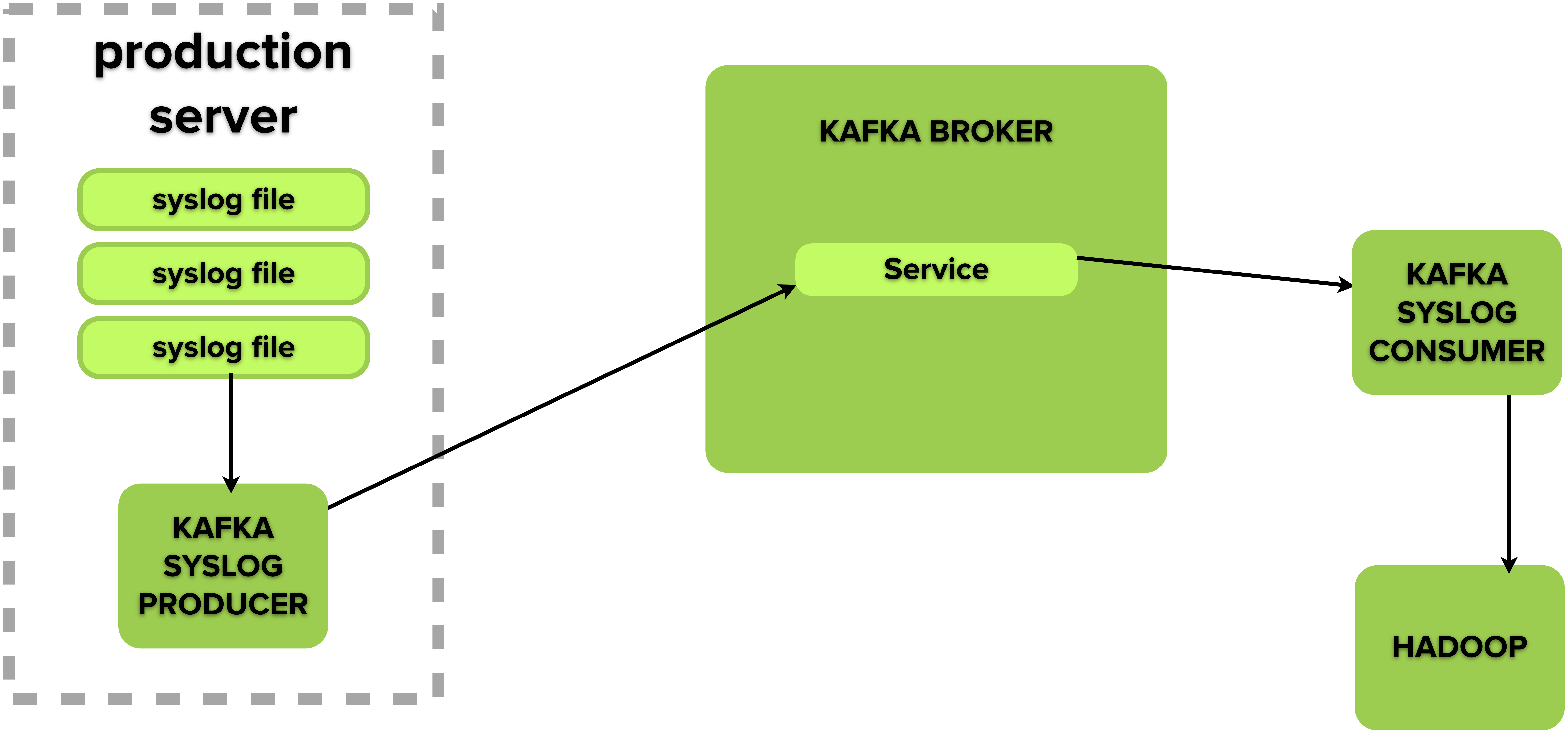
syslog file

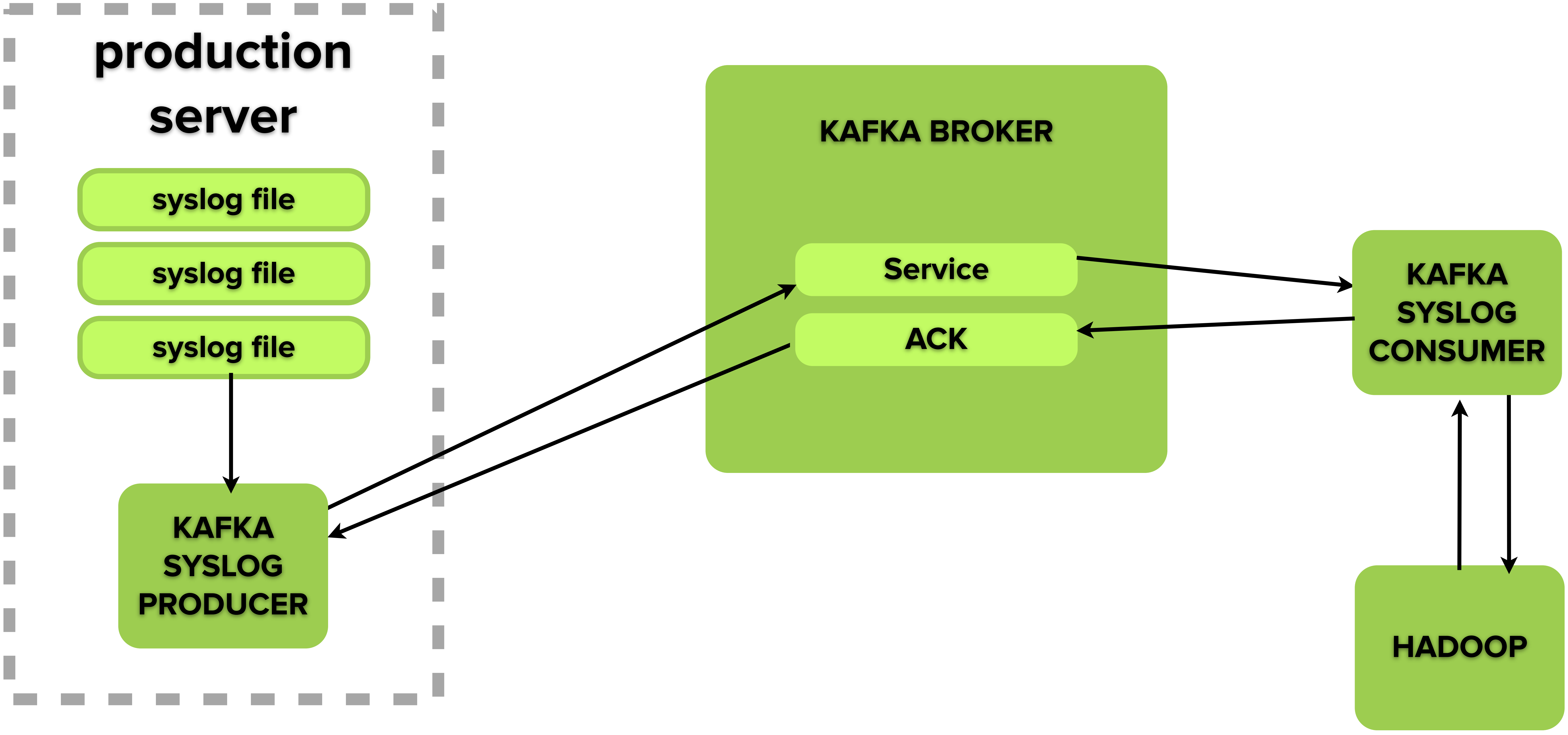
KAFKA SYSLOG PRODUCER

KAFKA BROKER

Service


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA BROKER
Service
KAFKA SYSLOG CONSUMER


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA BROKER
Service
KAFKA SYSLOG CONSUMER
HADOOP



production server

syslog file

syslog file

syslog file

KAFKA SYSLOG PRODUCER

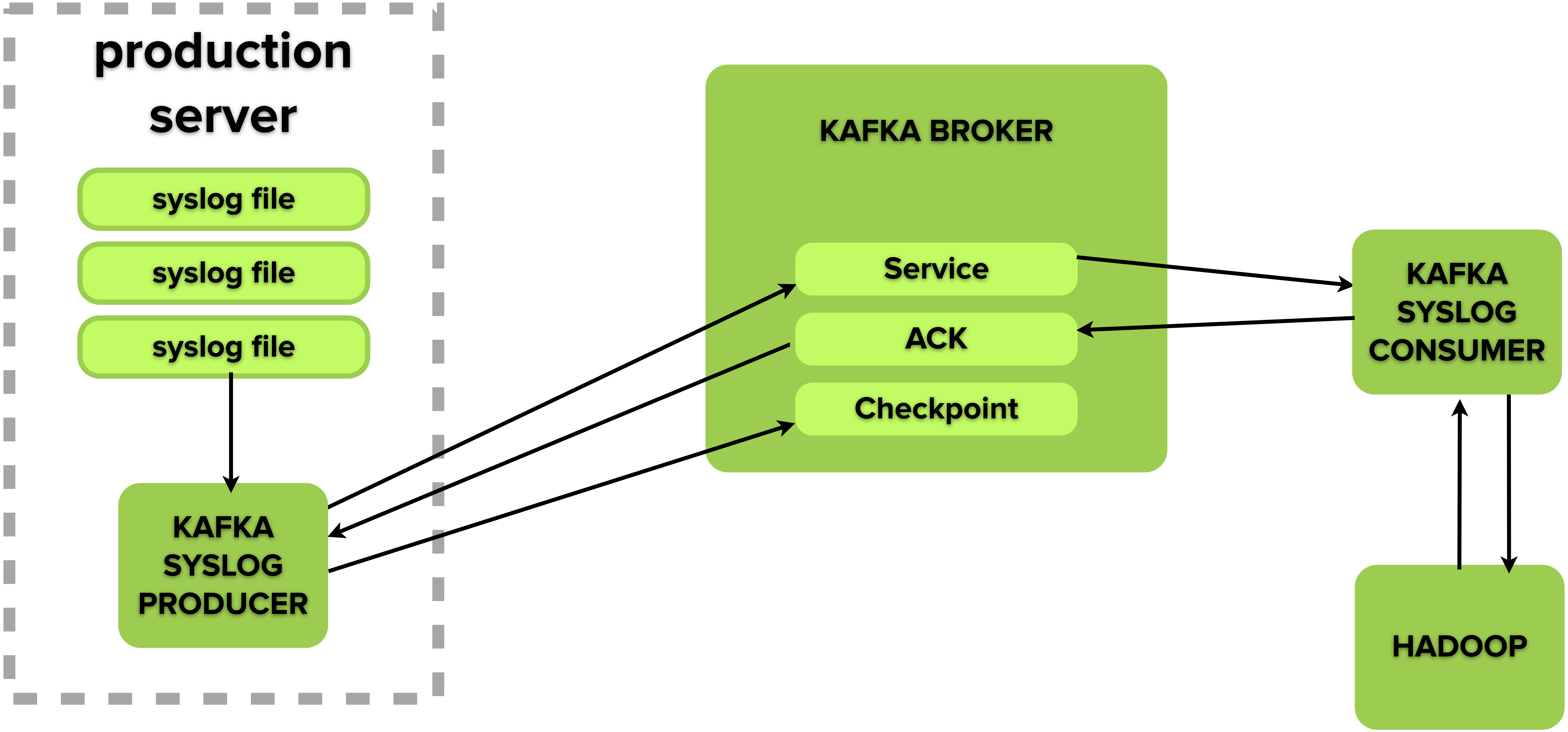
KAFKA BROKER

Service

ACK

KAFKA SYSLOG CONSUMER

HADOOP



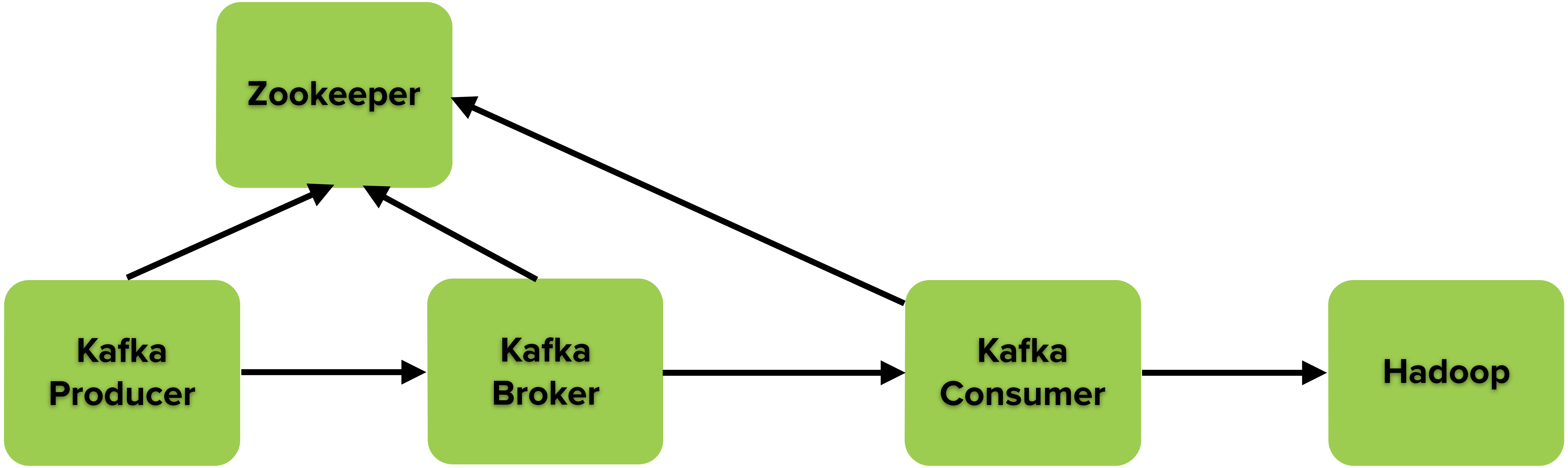
production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA BROKER
Service
ACK
Checkpoint
KAFKA SYSLOG CONSUMER
HADOOP

is that all?

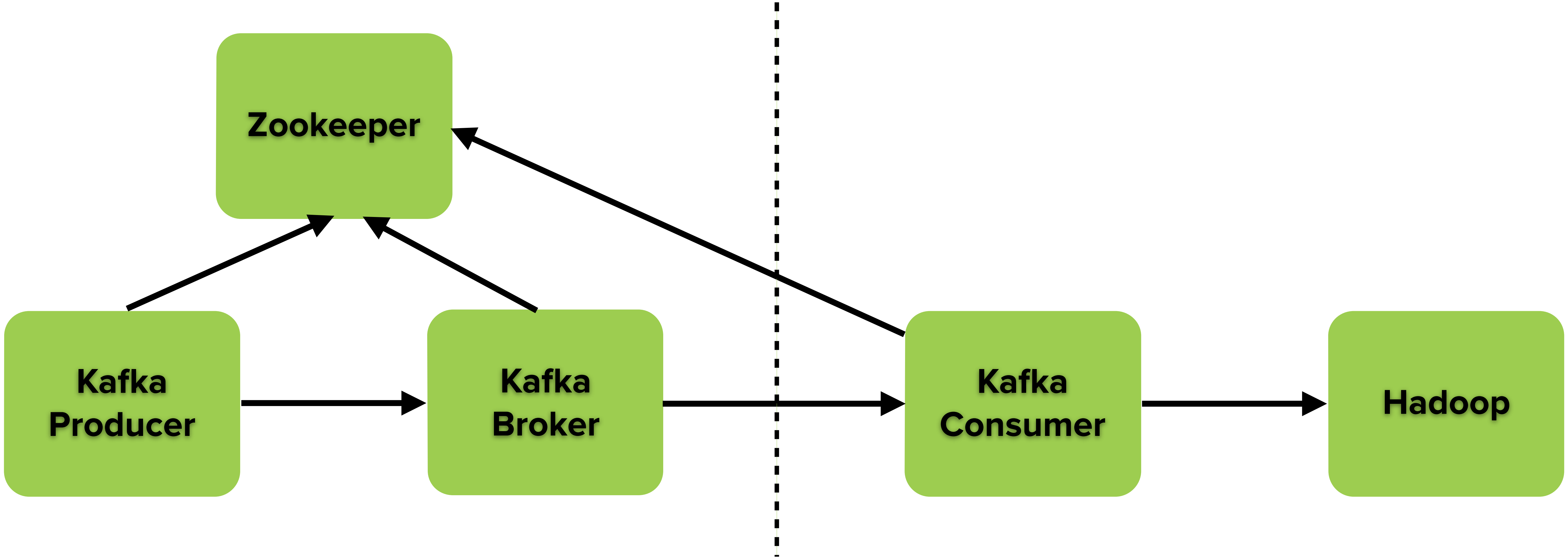
Piece of cake

right?

INSTALLED MY NEW KAFKA
IT DIDN'T WORK AS EXPECTED


Zookeeper
Kafka Producer
Kafka Broker
Kafka Consumer
Hadoop

Cross-site problems

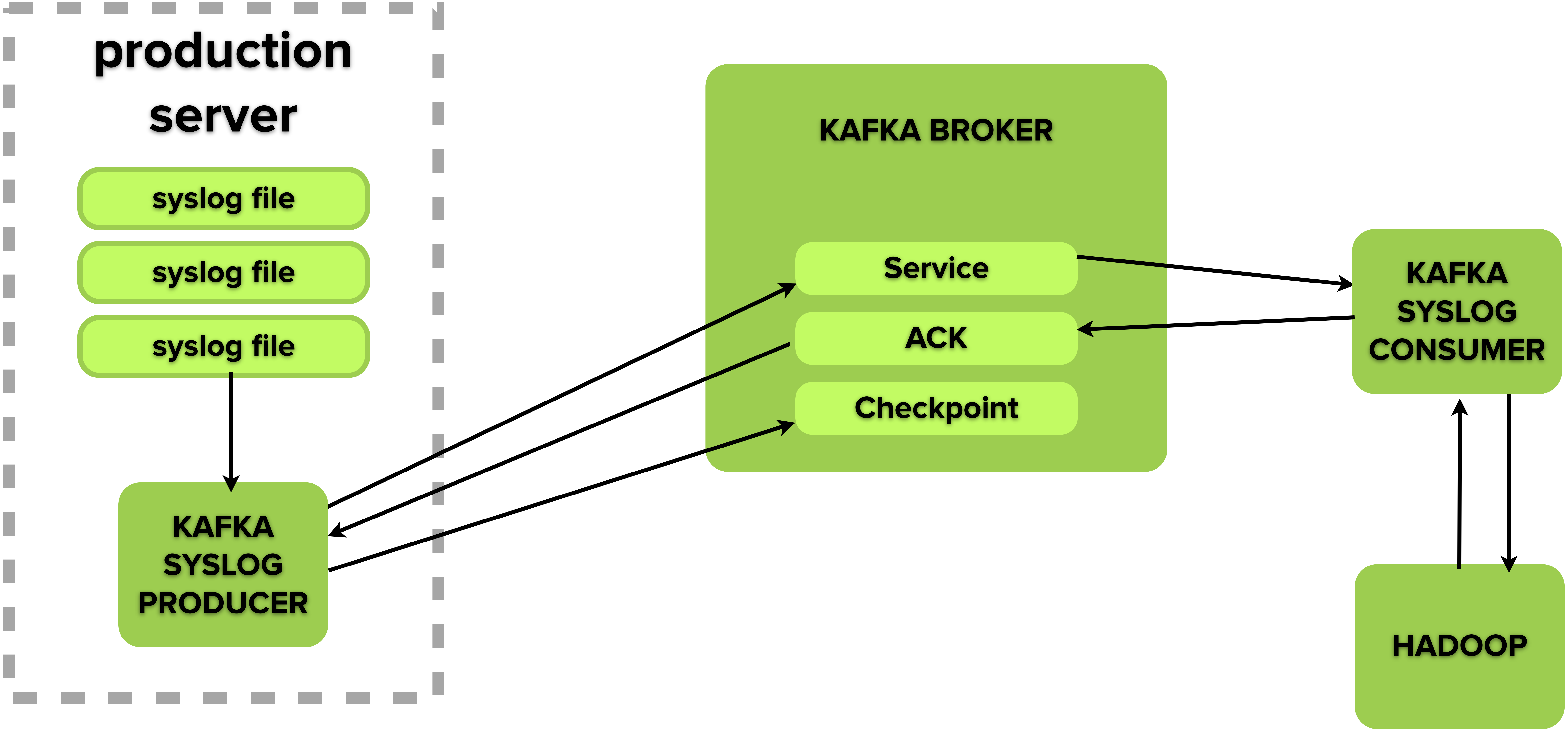


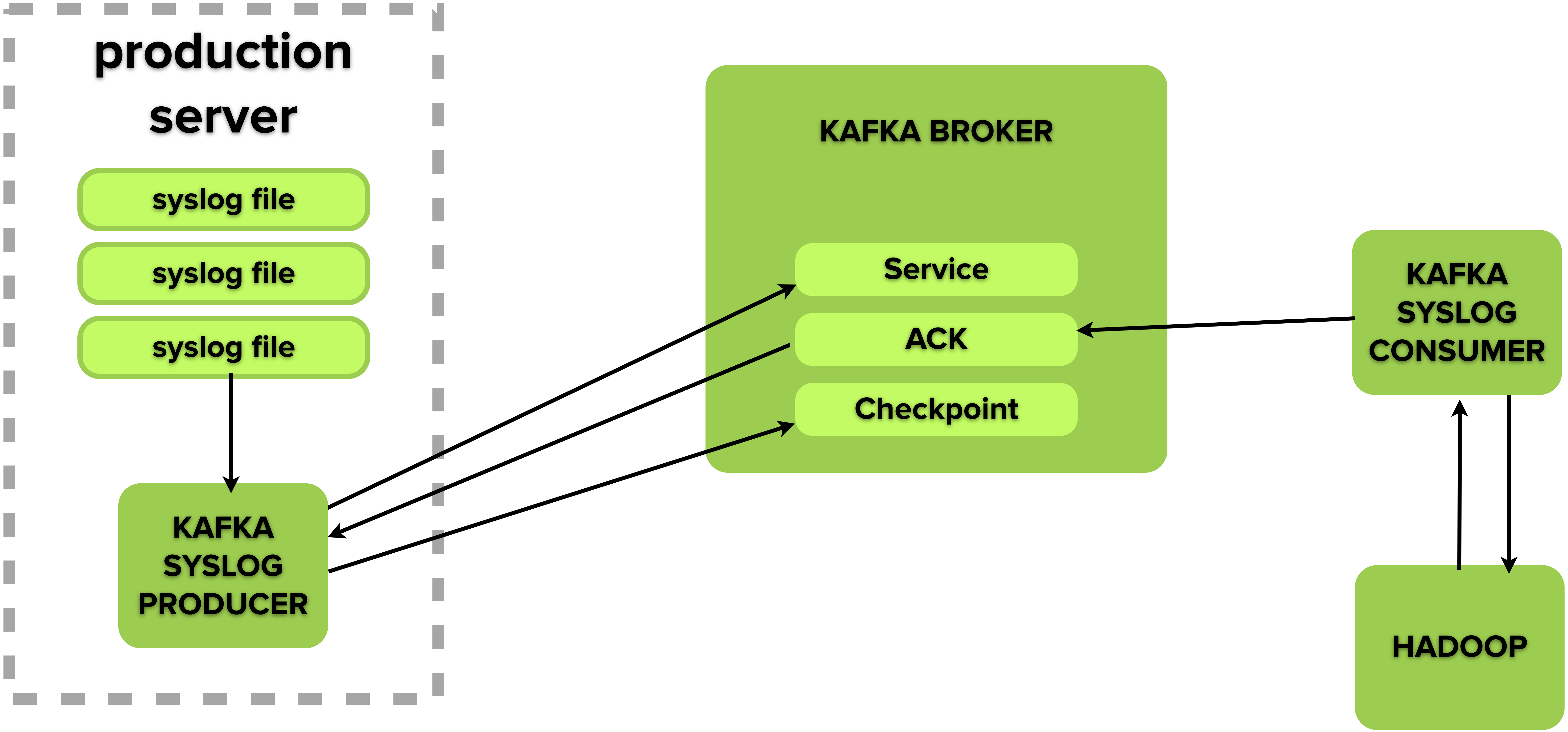
TCP window

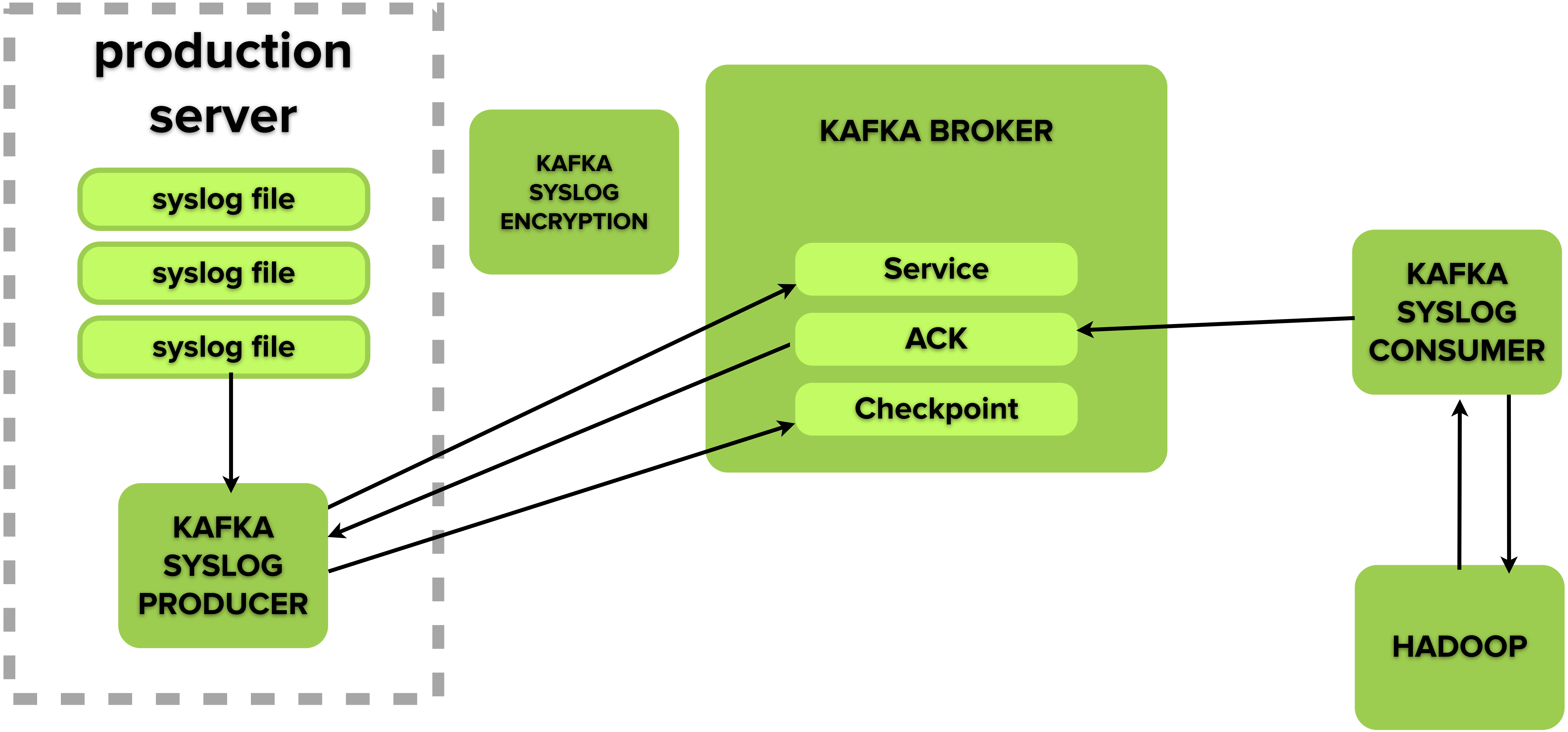
- TCP parameters for big latency
- linux TCP scaling algorithm

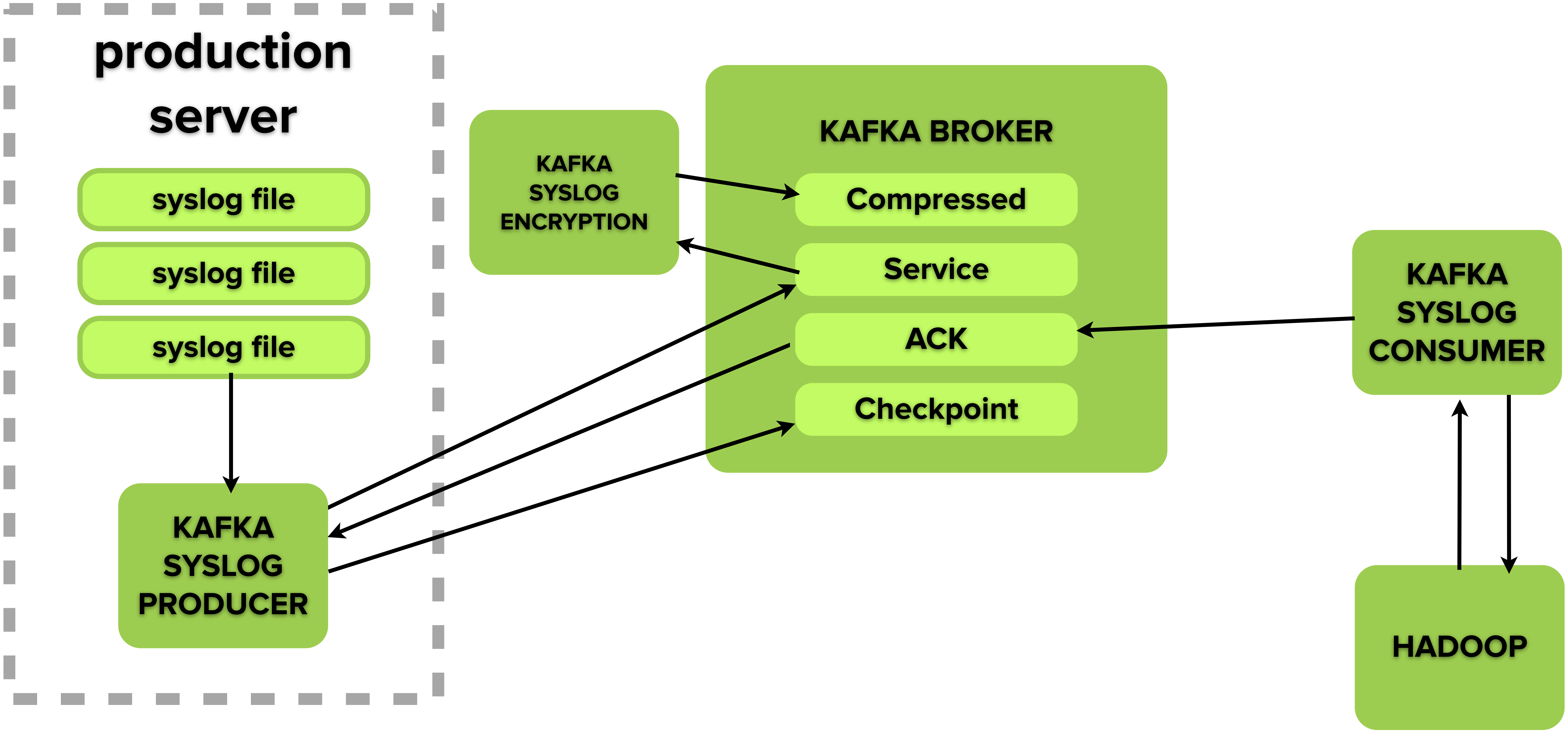
IPSEC

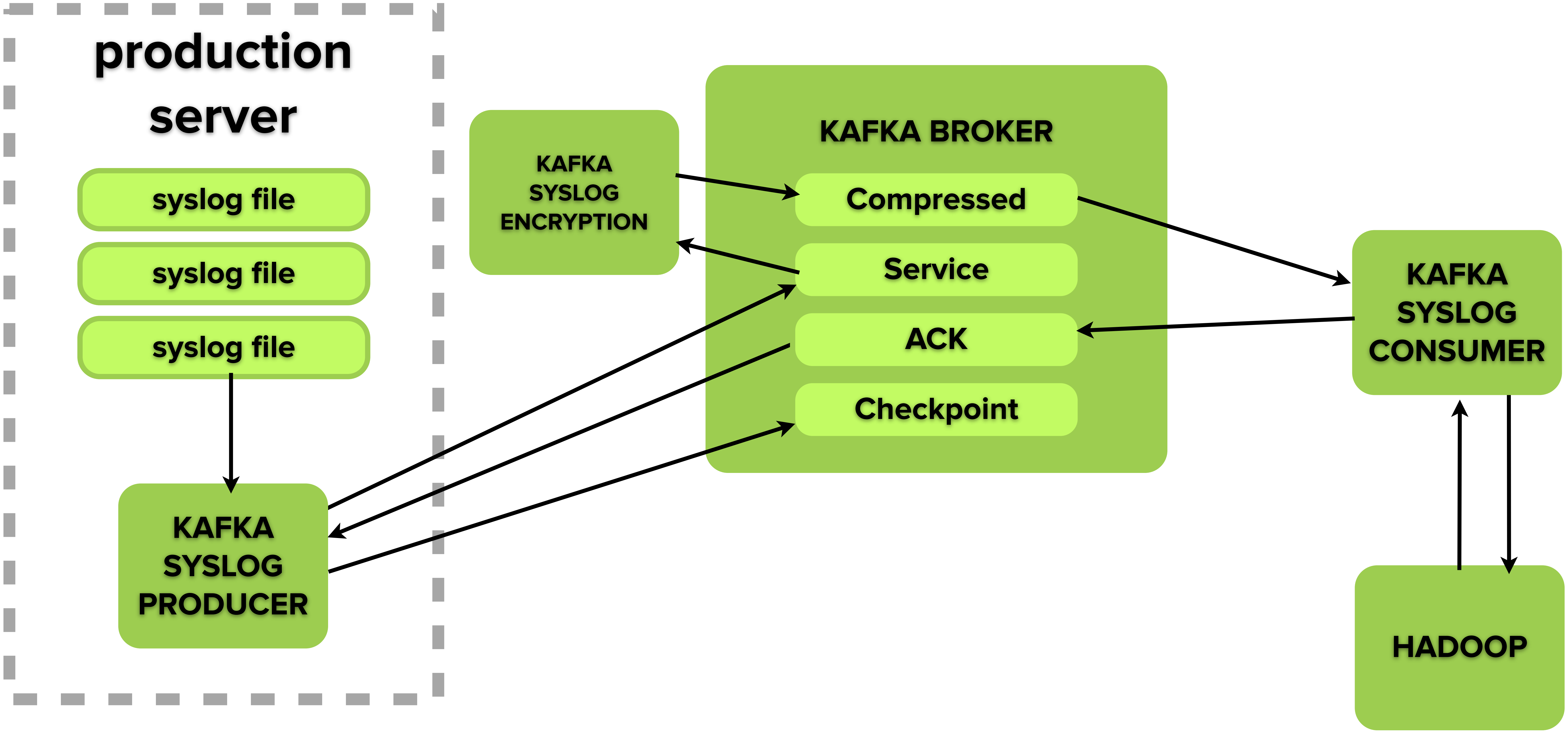
- linux IPSEC + firewall is slow
- major drop in throughput
- can not tweak it at app level


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA BROKER
Service
ACK
Checkpoint
KAFKA SYSLOG CONSUMER
HADOOP


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA BROKER
Service
ACK
Checkpoint
KAFKA SYSLOG CONSUMER
HADOOP


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA SYSLOG ENCRYPTION
KAFKA BROKER
Service
ACK
Checkpoint
KAFKA SYSLOG CONSUMER
HADOOP

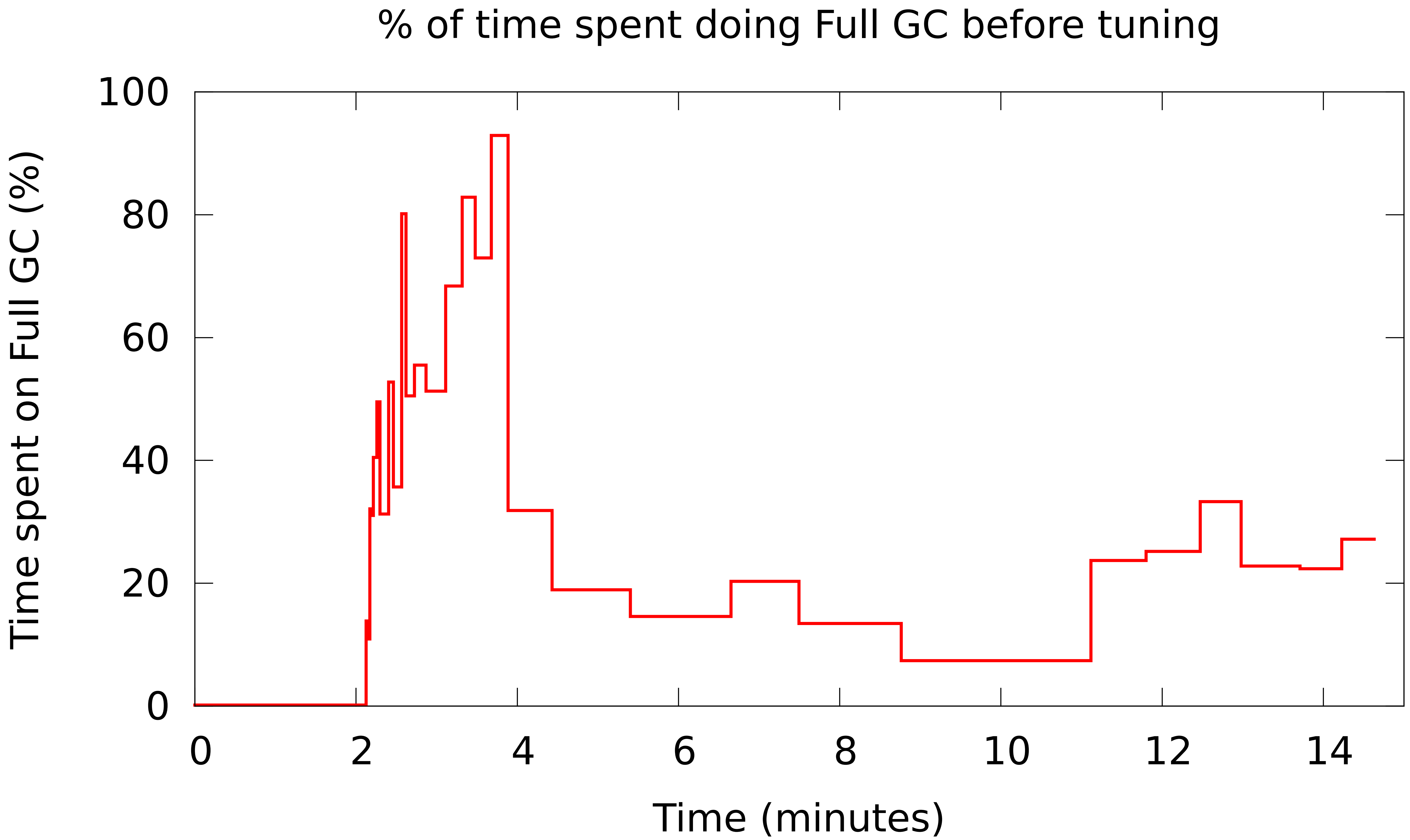

production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA SYSLOG ENCRYPTION
KAFKA BROKER
Compressed
Service
ACK
Checkpoint
KAFKA SYSLOG CONSUMER
HADOOP

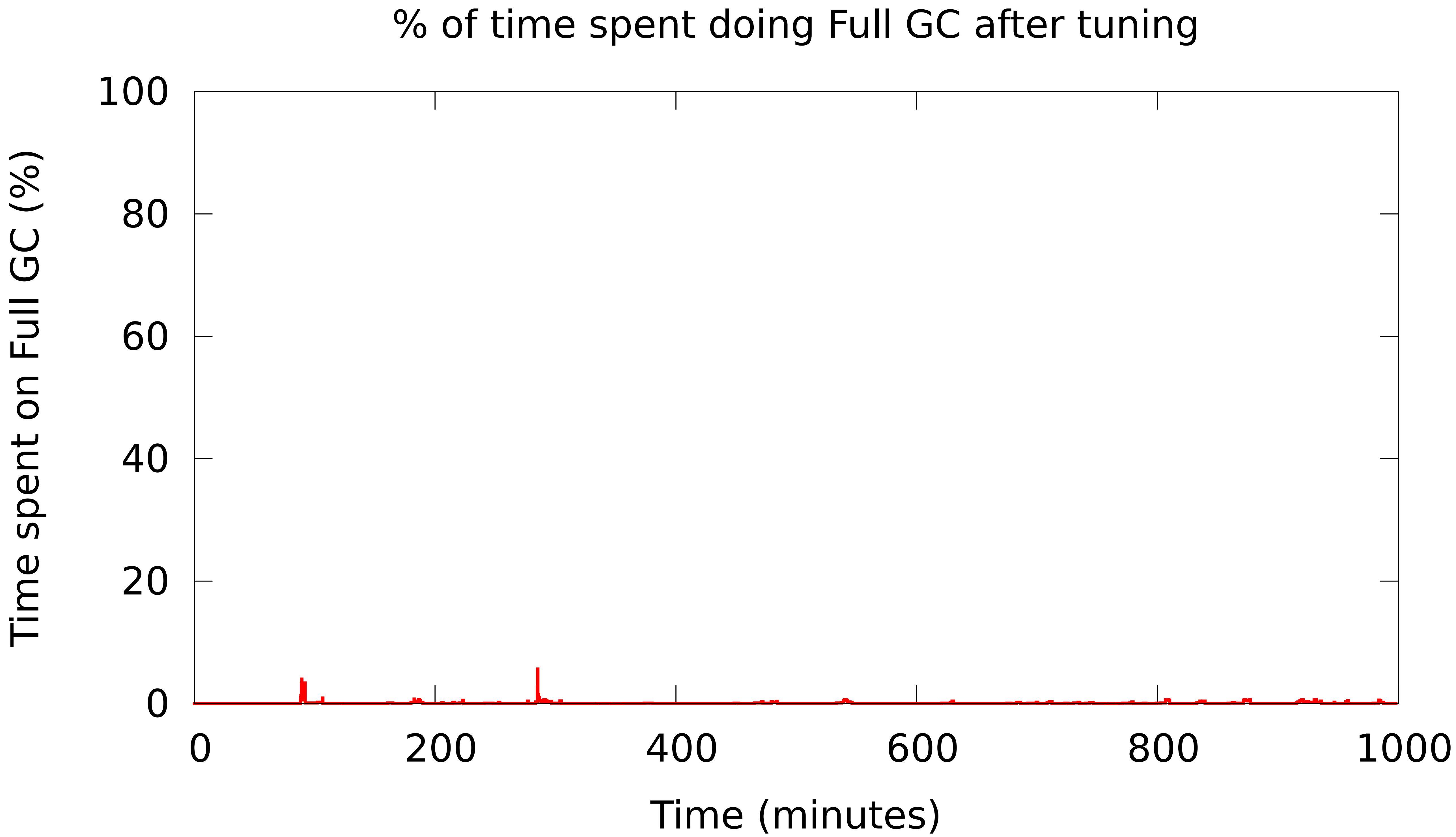

production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA SYSLOG ENCRYPTION
KAFKA BROKER
Compressed
Service
ACK
Checkpoint
KAFKA SYSLOG CONSUMER
HADOOP

I DON'T ALWAYS TEST MY CODE
BUT WHEN I DO I DO IT IN PRODUCTION
memegenerator.net

Garbage collector

- 50% of performance drop
- 25% of cpu time
- young generation tuning


% of time spent doing Full GC before tuning
Time spent on Full GC (%)
100
80
60
40
20
0
0
2
4
6
8
10
12
14
Time (minutes)


% of time spent doing Full GC after tuning
Time spent on Full GC (%)
100
80
60
40
20
0
0
200
400
600
800
1000
Time (minutes)

Hadoop replication factor

- stochastic failure mode
- no real ack from Hadoop
- files open for a long time

KEEP
CALM
AND
TRANSFER
DATA

Apache Storm

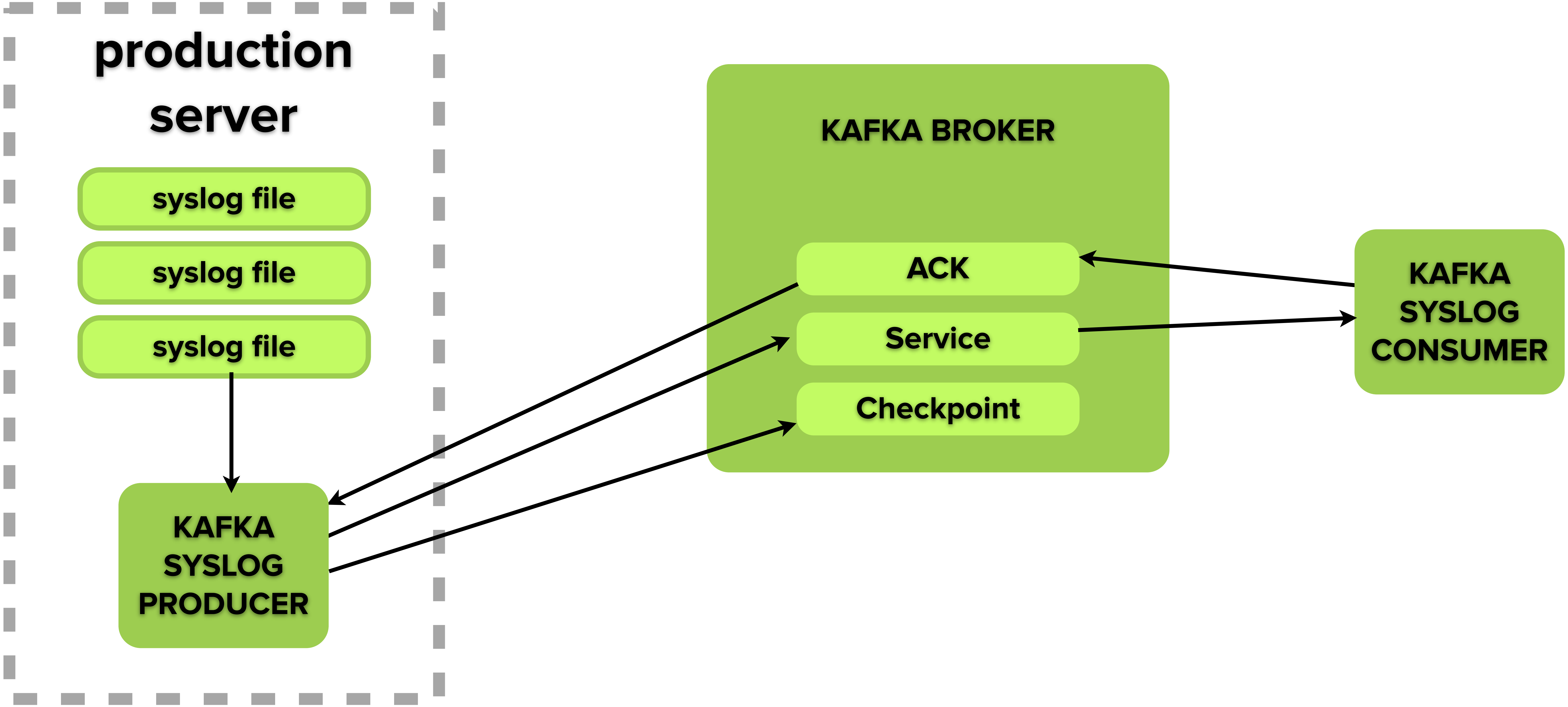
distributed computation framework

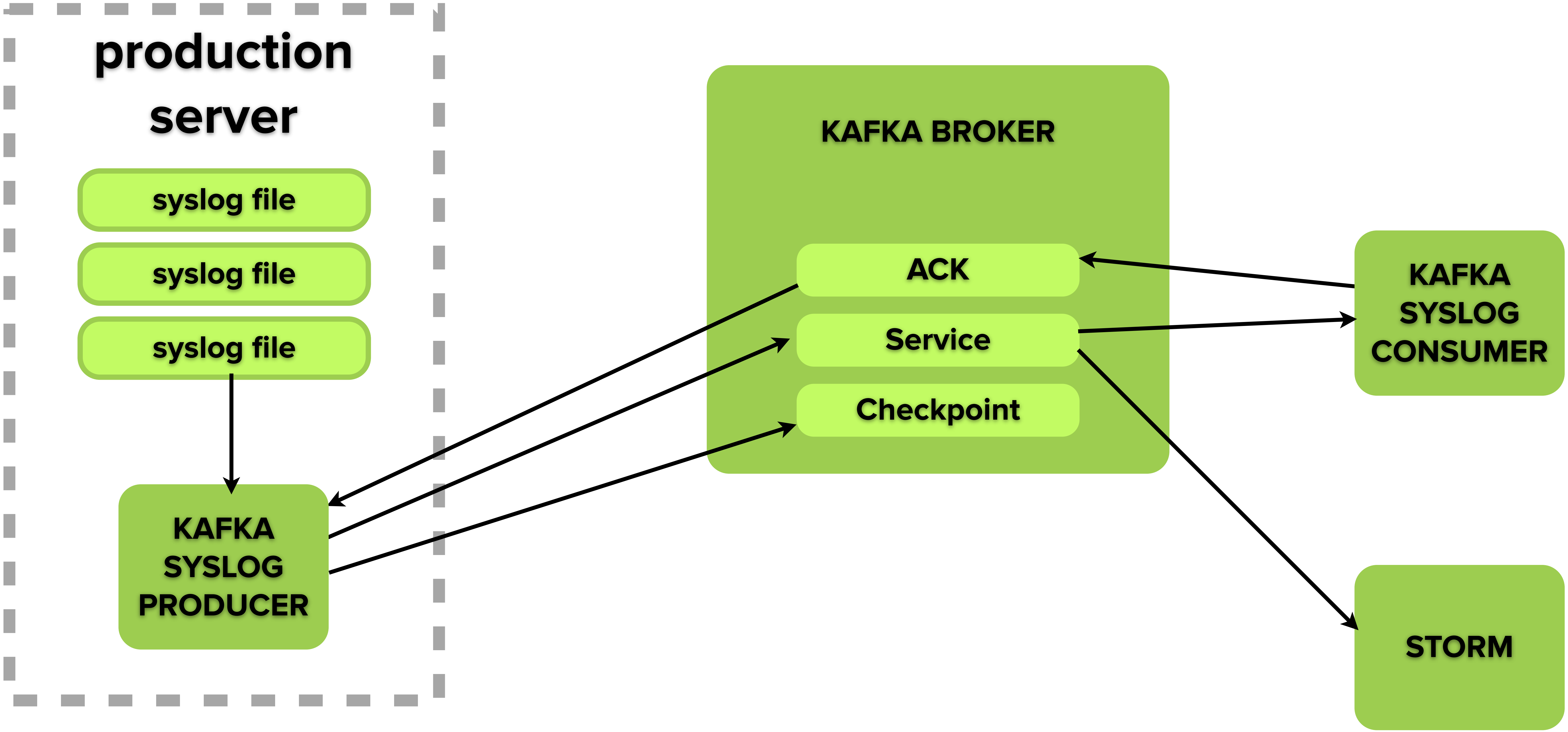
Storm

- abstractions: topology, bolt, stream, tuple, grouping
- great community
- ack + retries
- but not for reliable apps
  - use Hadoop instead

Kafka integration

- reliable data for reporting
- low latency data for RT


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA BROKER
ACK
Service
Checkpoint
KAFKA SYSLOG CONSUMER



production server

syslog file

syslog file

syslog file

KAFKA SYSLOG PRODUCER

KAFKA BROKER

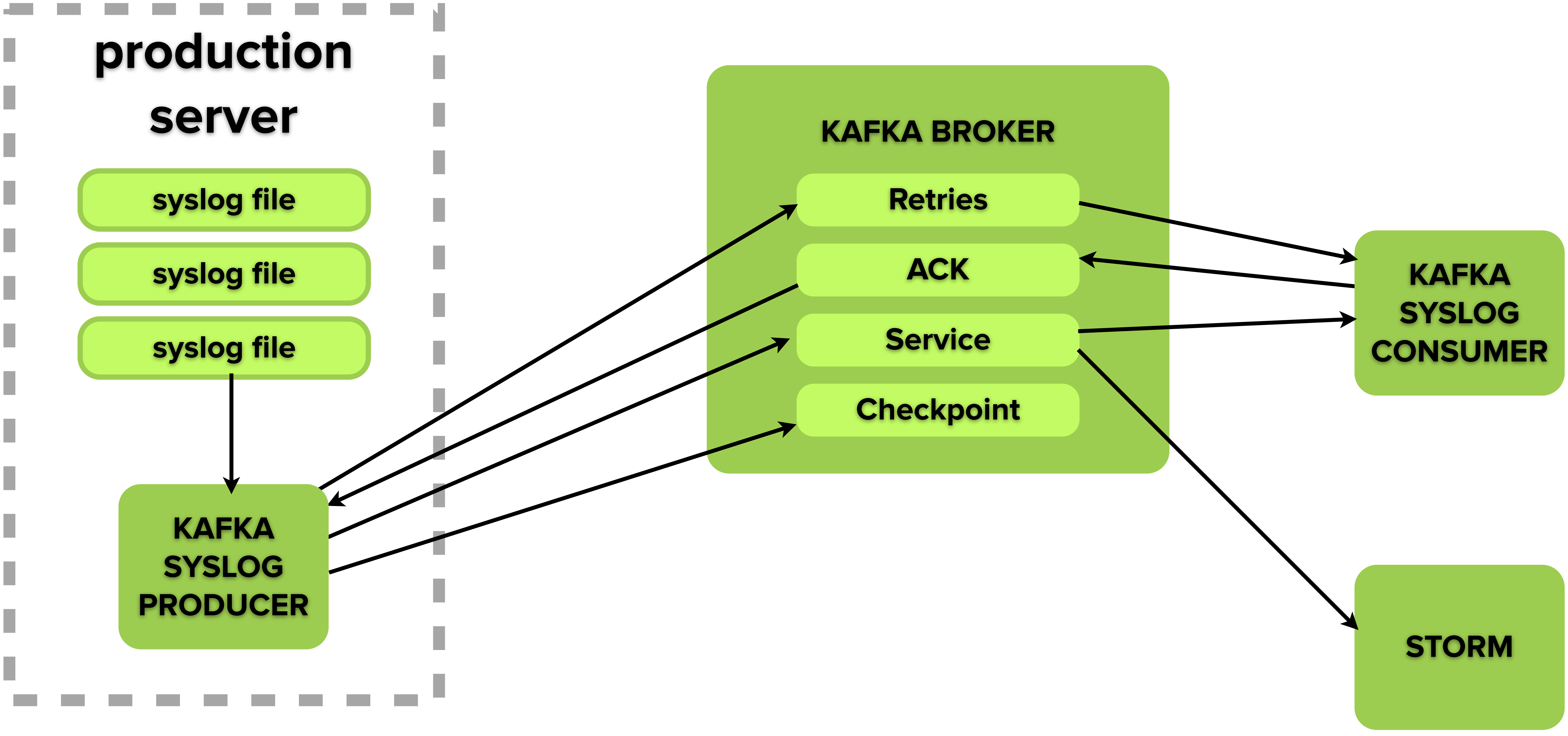
ACK

Service

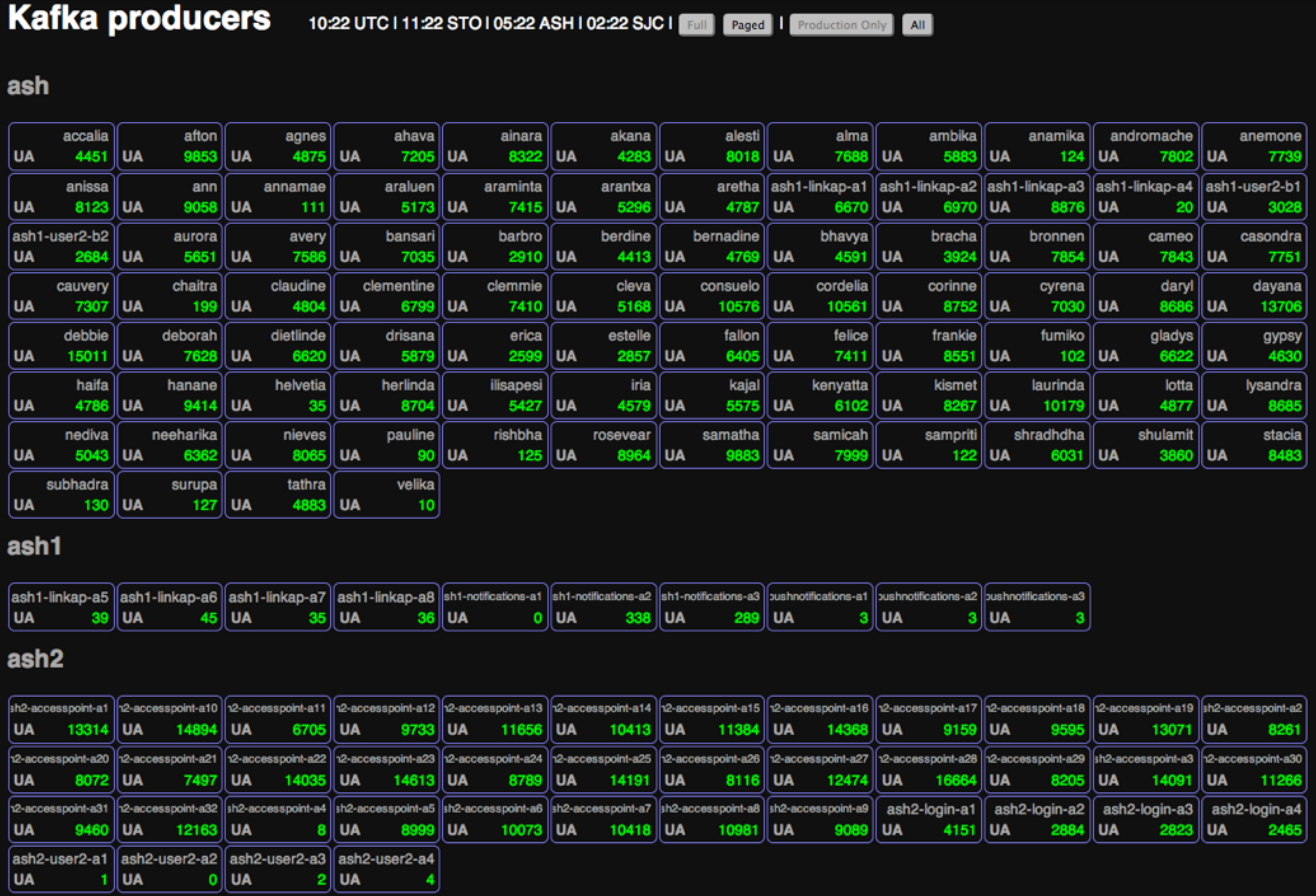
Checkpoint

KAFKA SYSLOG CONSUMER

STORM


production server
syslog file
syslog file
syslog file
KAFKA SYSLOG PRODUCER
KAFKA BROKER
Retries
ACK
Service
Checkpoint
KAFKA SYSLOG CONSUMER
STORM

RT apps



Kafka producers

10:22 UTC | 11:22 STO | 05:22 ASH | 02:22 SJC | Full Paged | Production Only All

ash

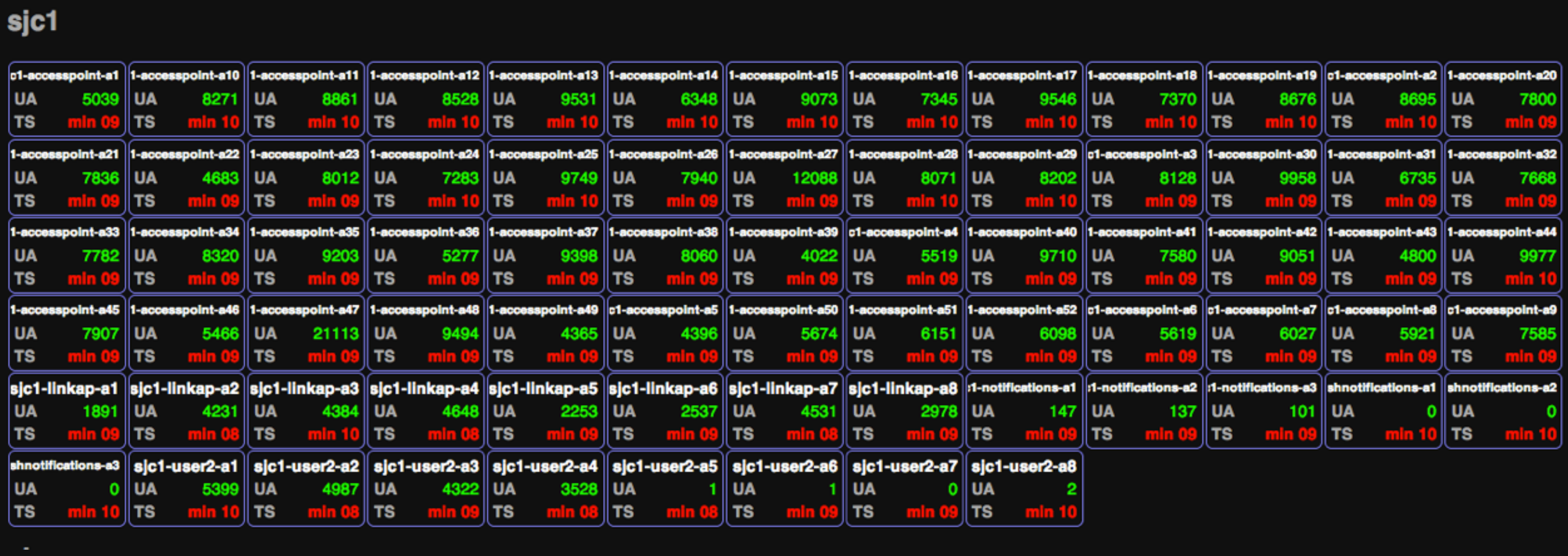
accalia UA 4451	afton UA 9853	agnes UA 4875	ahava UA 7205	ainara UA 8322	akana UA 4283	alesti UA 8018	alma UA 7688	ambika UA 5883	anamika UA 124	andromache UA 7802	anemone UA 7739
anissa UA 8123	ann UA 9058	annamae UA 111	araluen UA 5173	araminta UA 7415	arantxa UA 5296	aretha UA 4787	ash1-linkap-a1 UA 6670	ash1-linkap-a2 UA 6970	ash1-linkap-a3 UA 8876	ash1-linkap-a4 UA 20	ash1-user2-b1 UA 3028
ash1-user2-b2 UA 2684	aurora UA 5651	avery UA 7586	bansari UA 7035	barbro UA 2910	berdine UA 4413	bernadine UA 4769	bhavya UA 4591	bracha UA 3924	bronnen UA 7854	cameo UA 7843	casondra UA 7751
cauvery UA 7307	chaitra UA 199	claudine UA 4804	clementine UA 6799	clemmie UA 7410	cleva UA 5168	consuelo UA 10576	cordelia UA 10561	corinne UA 8752	cyrena UA 7030	daryl UA 8686	dayana UA 13706
debbie UA 15011	deborah UA 7628	dietlinde UA 6620	drisana UA 5879	erica UA 2599	estelle UA 2857	fallon UA 6405	felice UA 7411	frankie UA 8551	fumiko UA 102	gladys UA 6622	gypsy UA 4630
haifa UA 4786	hanane UA 9414	helvetia UA 35	herlinda UA 8704	ilisapesi UA 5427	iria UA 4579	kajal UA 5575	kenyatta UA 6102	kismet UA 8267	laurinda UA 10179	lotta UA 4877	lysandra UA 8685
nediva UA 5043	neeharika UA 6362	nieves UA 8065	pauline UA 90	rishbha UA 125	rosevear UA 8964	samatha UA 9883	samicah UA 7999	sampriti UA 122	shradhdha UA 6031	shulamit UA 3860	stacia UA 8483
subhadra UA 130	surupa UA 127	tathra UA 4883	velika UA 10								

ash1

ash1-linkap-a5 UA 39	ash1-linkap-a6 UA 45	ash1-linkap-a7 UA 35	ash1-linkap-a8 UA 36	sh1-notifications-a1 UA 0	sh1-notifications-a2 UA 338	sh1-notifications-a3 UA 289	ushnotifications-a1 UA 3	ushnotifications-a2 UA 3	ushnotifications-a3 UA 3

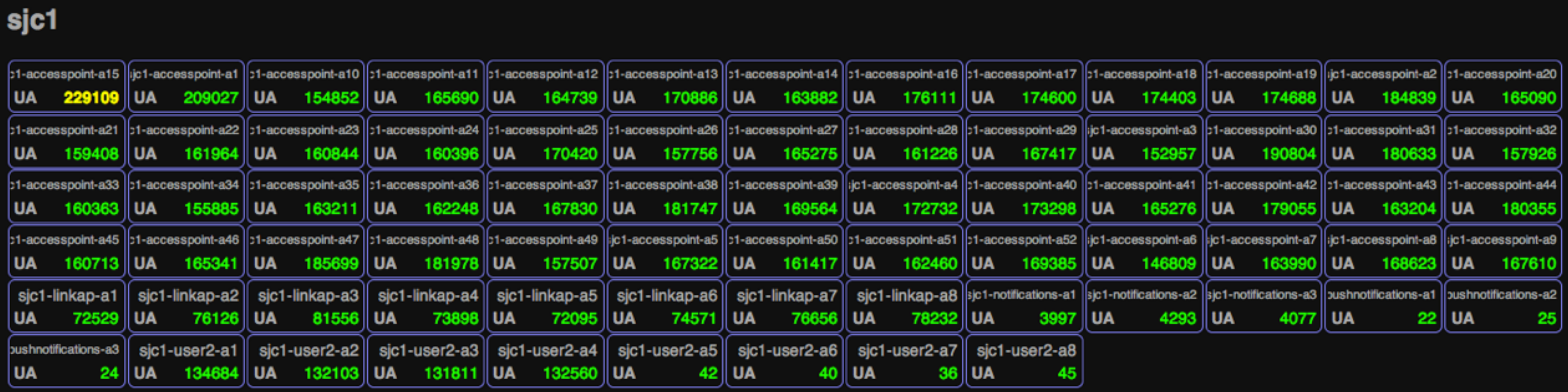
ash2

sh2-accesspoint-a1 UA 13314	2-accesspoint-a10 UA 14894	2-accesspoint-a11 UA 6705	2-accesspoint-a12 UA 9733	2-accesspoint-a13 UA 11656	2-accesspoint-a14 UA 10413	2-accesspoint-a15 UA 11384	2-accesspoint-a16 UA 14368	2-accesspoint-a17 UA 9159	2-accesspoint-a18 UA 9595	2-accesspoint-a19 UA 13071	sh2-accesspoint-a2 UA 8261
2-accesspoint-a20 UA 8072	2-accesspoint-a21 UA 7497	2-accesspoint-a22 UA 14035	2-accesspoint-a23 UA 14613	2-accesspoint-a24 UA 8789	2-accesspoint-a25 UA 14191	2-accesspoint-a26 UA 8116	2-accesspoint-a27 UA 12474	2-accesspoint-a28 UA 16664	2-accesspoint-a29 UA 8205	sh2-accesspoint-a3 UA 14091	2-accesspoint-a30 UA 11266
2-accesspoint-a31 UA 9460	2-accesspoint-a32 UA 12163	sh2-accesspoint-a4 UA 8	sh2-accesspoint-a5 UA 8999	sh2-accesspoint-a6 UA 10073	sh2-accesspoint-a7 UA 10418	sh2-accesspoint-a8 UA 10981	sh2-accesspoint-a9 UA 9089	ash2-login-a1 UA 4151	ash2-login-a2 UA 2884	ash2-login-a3 UA 2823	ash2-login-a4 UA 2465
ash2-user2-a1 UA 1	ash2-user2-a2 UA 0	ash2-user2-a3 UA 2	ash2-user2-a4 UA 4								



sjc1

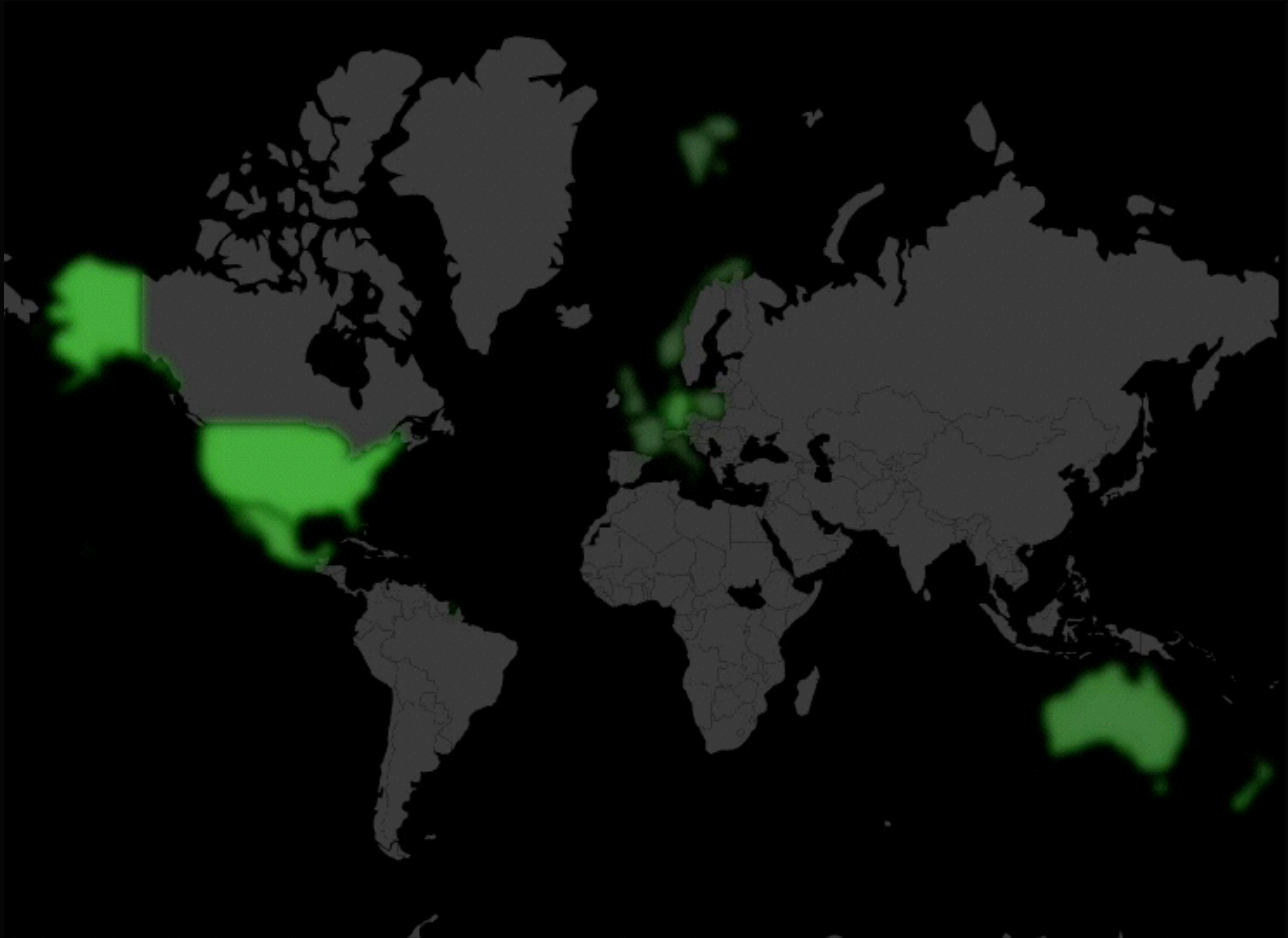
c1-accesspoint-a1 UA 5039 TS min 09	1-accesspoint-a10 UA 8271 TS min 10	1-accesspoint-a11 UA 8861 TS min 10	1-accesspoint-a12 UA 8528 TS min 10	1-accesspoint-a13 UA 9531 TS min 10	1-accesspoint-a14 UA 6348 TS min 10	1-accesspoint-a15 UA 9073 TS min 10	1-accesspoint-a16 UA 7345 TS min 10	1-accesspoint-a17 UA 9546 TS min 10	1-accesspoint-a18 UA 7370 TS min 10	1-accesspoint-a19 UA 8676 TS min 10	c1-accesspoint-a2 UA 8695 TS min 10	1-accesspoint-a20 UA 7800 TS min 09
1-accesspoint-a21 UA 7836 TS min 09	1-accesspoint-a22 UA 4683 TS min 09	1-accesspoint-a23 UA 8012 TS min 09	1-accesspoint-a24 UA 7283 TS min 10	1-accesspoint-a25 UA 9749 TS min 10	1-accesspoint-a26 UA 7940 TS min 09	1-accesspoint-a27 UA 12088 TS min 09	1-accesspoint-a28 UA 8071 TS min 10	1-accesspoint-a29 UA 8202 TS min 10	c1-accesspoint-a3 UA 8128 TS min 09	1-accesspoint-a30 UA 9958 TS min 09	1-accesspoint-a31 UA 6735 TS min 09	1-accesspoint-a32 UA 7668 TS min 09
1-accesspoint-a33 UA 7782 TS min 09	1-accesspoint-a34 UA 8320 TS min 09	1-accesspoint-a35 UA 9203 TS min 09	1-accesspoint-a36 UA 5277 TS min 09	1-accesspoint-a37 UA 9398 TS min 09	1-accesspoint-a38 UA 8060 TS min 09	1-accesspoint-a39 UA 4022 TS min 09	c1-accesspoint-a4 UA 5519 TS min 09	1-accesspoint-a40 UA 9710 TS min 09	1-accesspoint-a41 UA 7580 TS min 09	1-accesspoint-a42 UA 9051 TS min 09	1-accesspoint-a43 UA 4800 TS min 09	1-accesspoint-a44 UA 9977 TS min 10
1-accesspoint-a45 UA 7907 TS min 09	1-accesspoint-a46 UA 5466 TS min 09	1-accesspoint-a47 UA 21113 TS min 09	1-accesspoint-a48 UA 9494 TS min 09	1-accesspoint-a49 UA 4365 TS min 09	c1-accesspoint-a5 UA 4396 TS min 09	1-accesspoint-a50 UA 5674 TS min 09	1-accesspoint-a51 UA 6151 TS min 09	1-accesspoint-a52 UA 6098 TS min 09	c1-accesspoint-a6 UA 5619 TS min 09	c1-accesspoint-a7 UA 6027 TS min 09	c1-accesspoint-a8 UA 5921 TS min 09	c1-accesspoint-a9 UA 7585 TS min 09
sjc1-linkap-a1 UA 1891 TS min 09	sjc1-linkap-a2 UA 4231 TS min 08	sjc1-linkap-a3 UA 4384 TS min 10	sjc1-linkap-a4 UA 4648 TS min 08	sjc1-linkap-a5 UA 2253 TS min 09	sjc1-linkap-a6 UA 2537 TS min 09	sjc1-linkap-a7 UA 4531 TS min 08	sjc1-linkap-a8 UA 2978 TS min 09	:1-notifications-a1 UA 147 TS min 09	:1-notifications-a2 UA 137 TS min 09	:1-notifications-a3 UA 101 TS min 09	shnotifications-a1 UA 0 TS min 10	shnotifications-a2 UA 0 TS min 10
shnotifications-a3 UA 0 TS min 10	sjc1-user2-a1 UA 5399 TS min 10	sjc1-user2-a2 UA 4987 TS min 08	sjc1-user2-a3 UA 4322 TS min 09	sjc1-user2-a4 UA 3528 TS min 08	sjc1-user2-a5 UA 1 TS min 08	sjc1-user2-a6 UA 1 TS min 09	sjc1-user2-a7 UA 0 TS min 09	sjc1-user2-a8 UA 2 TS min 10				

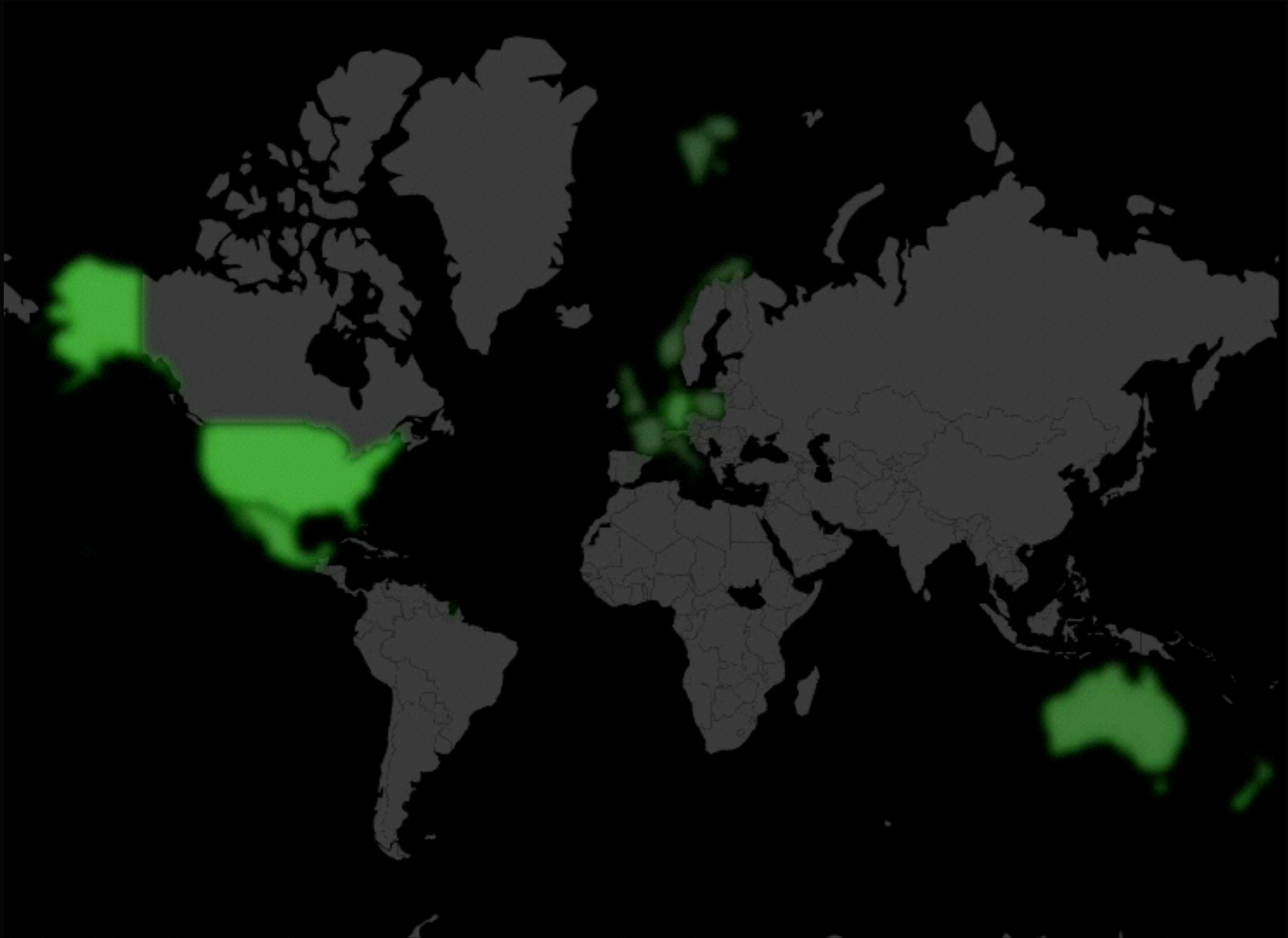


sjc1

c1-accesspoint-a15 UA 229109	jc1-accesspoint-a1 UA 209027	c1-accesspoint-a10 UA 154852	c1-accesspoint-a11 UA 165690	c1-accesspoint-a12 UA 164739	c1-accesspoint-a13 UA 170886	c1-accesspoint-a14 UA 163882	c1-accesspoint-a16 UA 176111	c1-accesspoint-a17 UA 174600	c1-accesspoint-a18 UA 174403	c1-accesspoint-a19 UA 174688	jc1-accesspoint-a2 UA 184839	c1-accesspoint-a20 UA 165090
c1-accesspoint-a21 UA 159408	c1-accesspoint-a22 UA 161964	c1-accesspoint-a23 UA 160844	c1-accesspoint-a24 UA 160396	c1-accesspoint-a25 UA 170420	c1-accesspoint-a26 UA 157756	c1-accesspoint-a27 UA 165275	c1-accesspoint-a28 UA 161226	c1-accesspoint-a29 UA 167417	jc1-accesspoint-a3 UA 152957	c1-accesspoint-a30 UA 190804	c1-accesspoint-a31 UA 180633	c1-accesspoint-a32 UA 157926
c1-accesspoint-a33 UA 160363	c1-accesspoint-a34 UA 155885	c1-accesspoint-a35 UA 163211	c1-accesspoint-a36 UA 162248	c1-accesspoint-a37 UA 167830	c1-accesspoint-a38 UA 181747	c1-accesspoint-a39 UA 169564	jc1-accesspoint-a4 UA 172732	c1-accesspoint-a40 UA 173298	c1-accesspoint-a41 UA 165276	c1-accesspoint-a42 UA 179055	c1-accesspoint-a43 UA 163204	c1-accesspoint-a44 UA 180355
c1-accesspoint-a45 UA 160713	c1-accesspoint-a46 UA 165341	c1-accesspoint-a47 UA 185699	c1-accesspoint-a48 UA 181978	c1-accesspoint-a49 UA 157507	jc1-accesspoint-a5 UA 167322	c1-accesspoint-a50 UA 161417	c1-accesspoint-a51 UA 162460	c1-accesspoint-a52 UA 169385	jc1-accesspoint-a6 UA 146809	jc1-accesspoint-a7 UA 163990	jc1-accesspoint-a8 UA 168623	jc1-accesspoint-a9 UA 167610
sjc1-linkap-a1 UA 72529	sjc1-linkap-a2 UA 76126	sjc1-linkap-a3 UA 81556	sjc1-linkap-a4 UA 73898	sjc1-linkap-a5 UA 72095	sjc1-linkap-a6 UA 74571	sjc1-linkap-a7 UA 76656	sjc1-linkap-a8 UA 78232	jc1-notifications-a1 UA 3997	jc1-notifications-a2 UA 4293	jc1-notifications-a3 UA 4077	ushnotifications-a1 UA 22	ushnotifications-a2 UA 25
ushnotifications-a3 UA 24	sjc1-user2-a1 UA 134684	sjc1-user2-a2 UA 132103	sjc1-user2-a3 UA 131811	sjc1-user2-a4 UA 132560	sjc1-user2-a5 UA 42	sjc1-user2-a6 UA 40	sjc1-user2-a7 UA 36	sjc1-user2-a8 UA 45				

Faroe Islands
Norway
Bergen
Turk
Baltic Sea
North Sea
Glasgow
United Kingdom
Isle of Ma
Leeds
erpool
Dublin
Ireland
Birmingham
Bristol
London
Celtic Sea
Copenhagen
De
Gdansk
Bremen
Amste
Netherlands
Berlin
Poland
Poznan
Warsaw
Wroclaw
Lodz
Belgium
Colo
Germany
Frankfurt
Luxembourg
Prague
Czech Republic
Krakow
Paris
Slovakia
Vienna
Bratislava
Zurich
Austria
Budapest
Hungary
Switzerland
France
Bay of Biscay
Lyon
Slovenia
Zagreb
Milan
Torino
Croatia
Bosnia and Herzegovina
Serbia
Toulouse





Netherlands

Netherlands

Thanks!

Pablo Barrera <pablo@spotify.com>

Want to join the band?
spotify.com/jobs