FROM RESEARCH TO INDUSTRY

Lustre-HSM

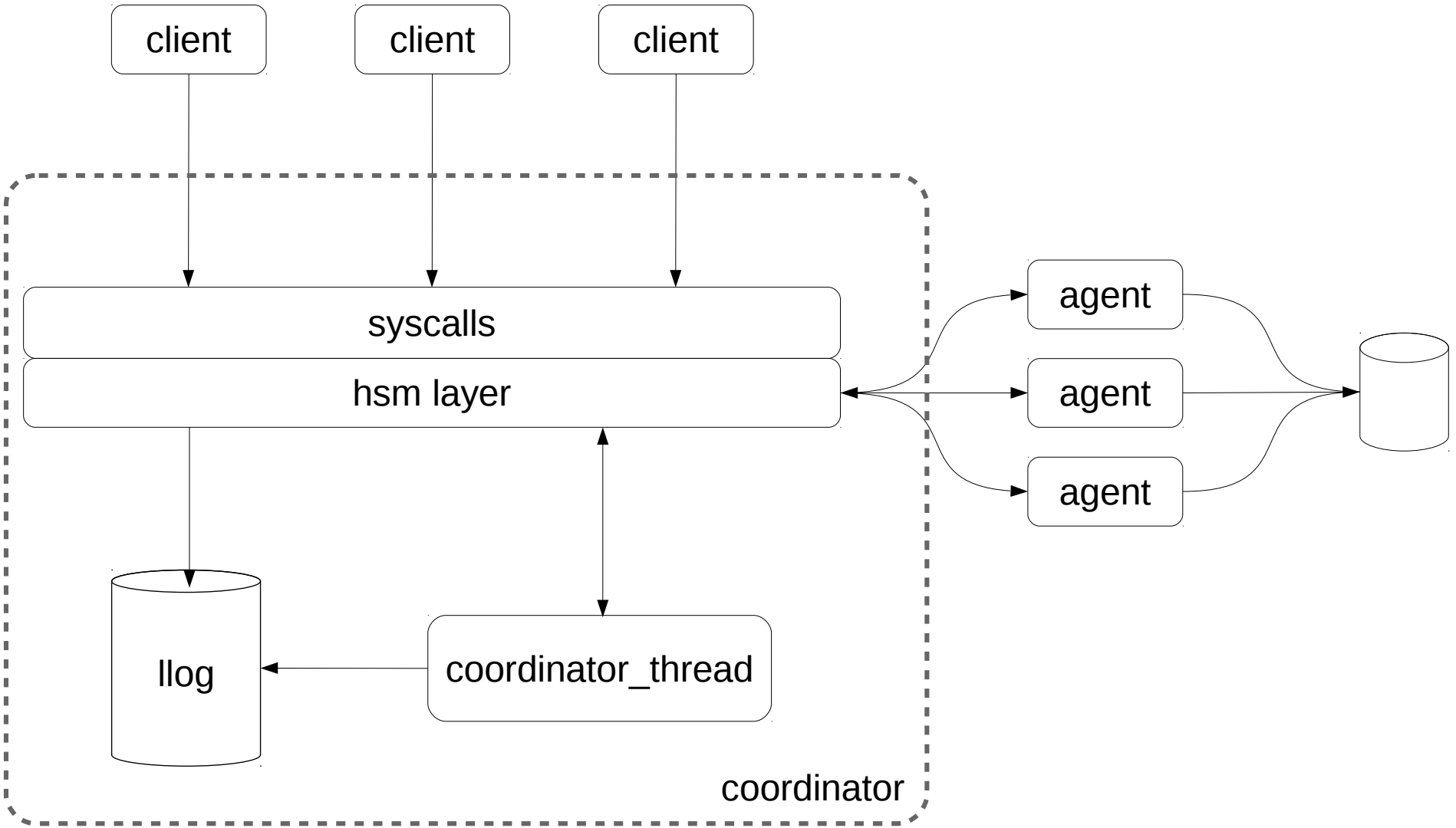
Progress report on Lustre HSM and proposal for a new design of the coordinator

www.cea.fr

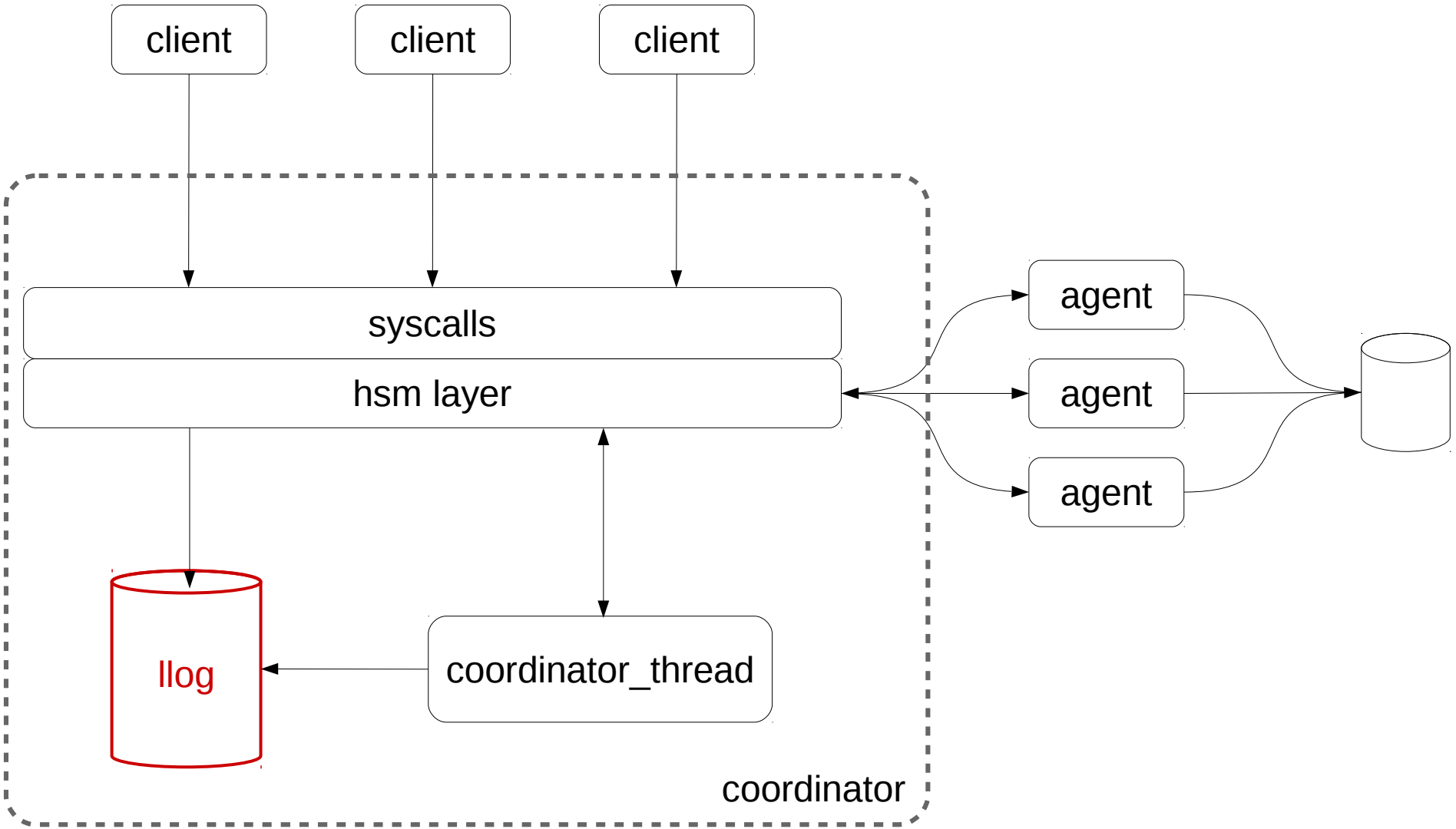
October 6, 2017

Quentin Bouget

Introduction: coordinator's design



Contended design



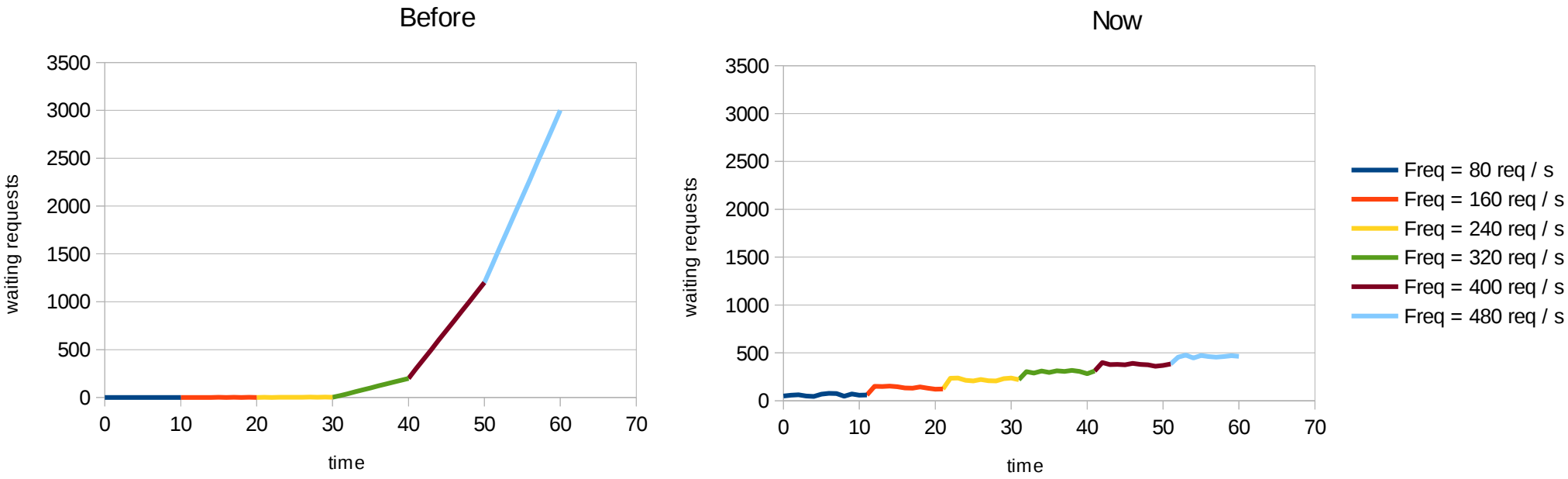
History (~6 months)

bugfix / **performance** / **feature**

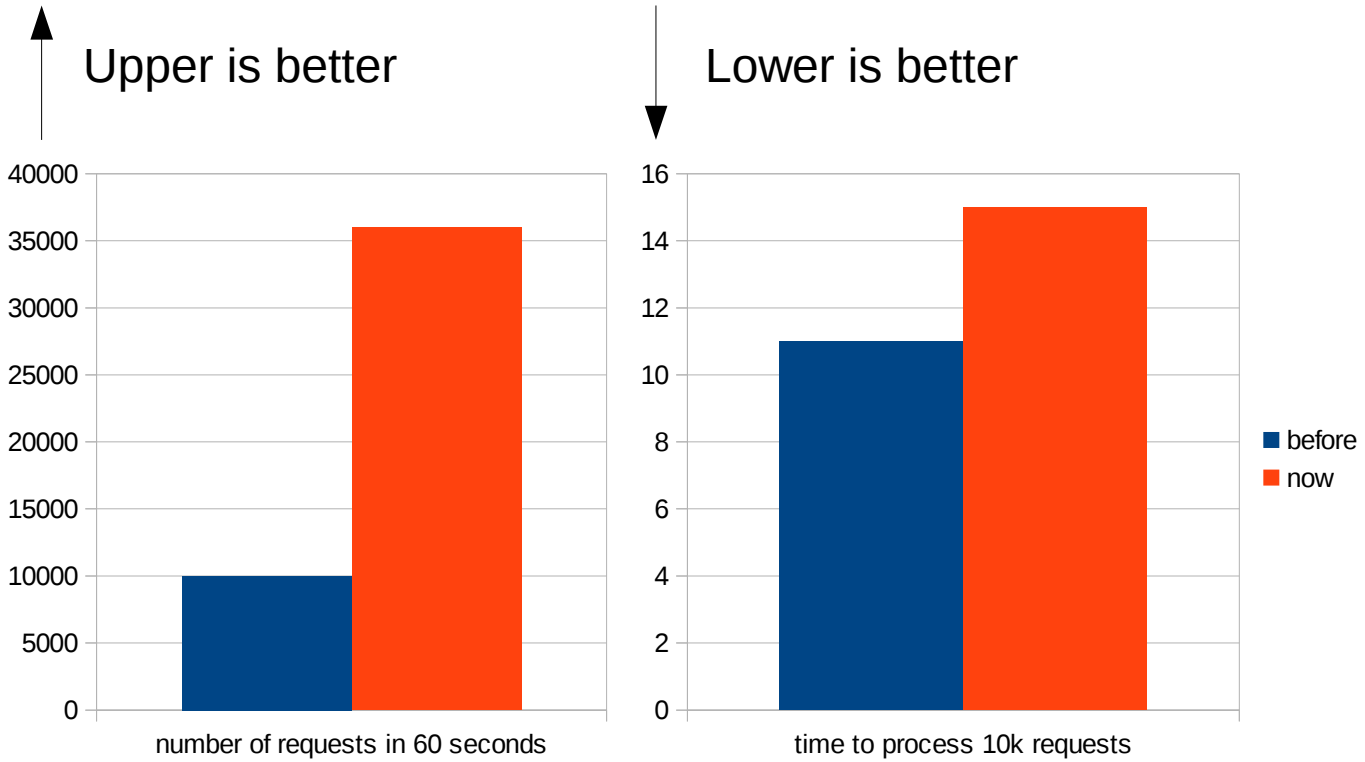
- **f3a4152** LU-7988 hsm: update many cookie status at once
- **afc9ff6** LU-7988 hsm: added coordinator housekeeping flag
- **cc6ef11** LU-7988 hsm: run HSM coordinator once per second at most
- **f11a502** LU-7988 hsm: mark the cdt as stopped when its thread exits
- **dd4b034** LU-4640 mdt: implement Remove Archive on Last Unlink policy
- **65effa6** LU-9338 hsm: cache agent record locations
- **2d5babe** LU-9312 hsm: convert cdt_llog_lock to a rw semaphore
- **958198e** LU-7988 hsm: change coordinator start/stop mechanisms

Performance improvements

Lower is better



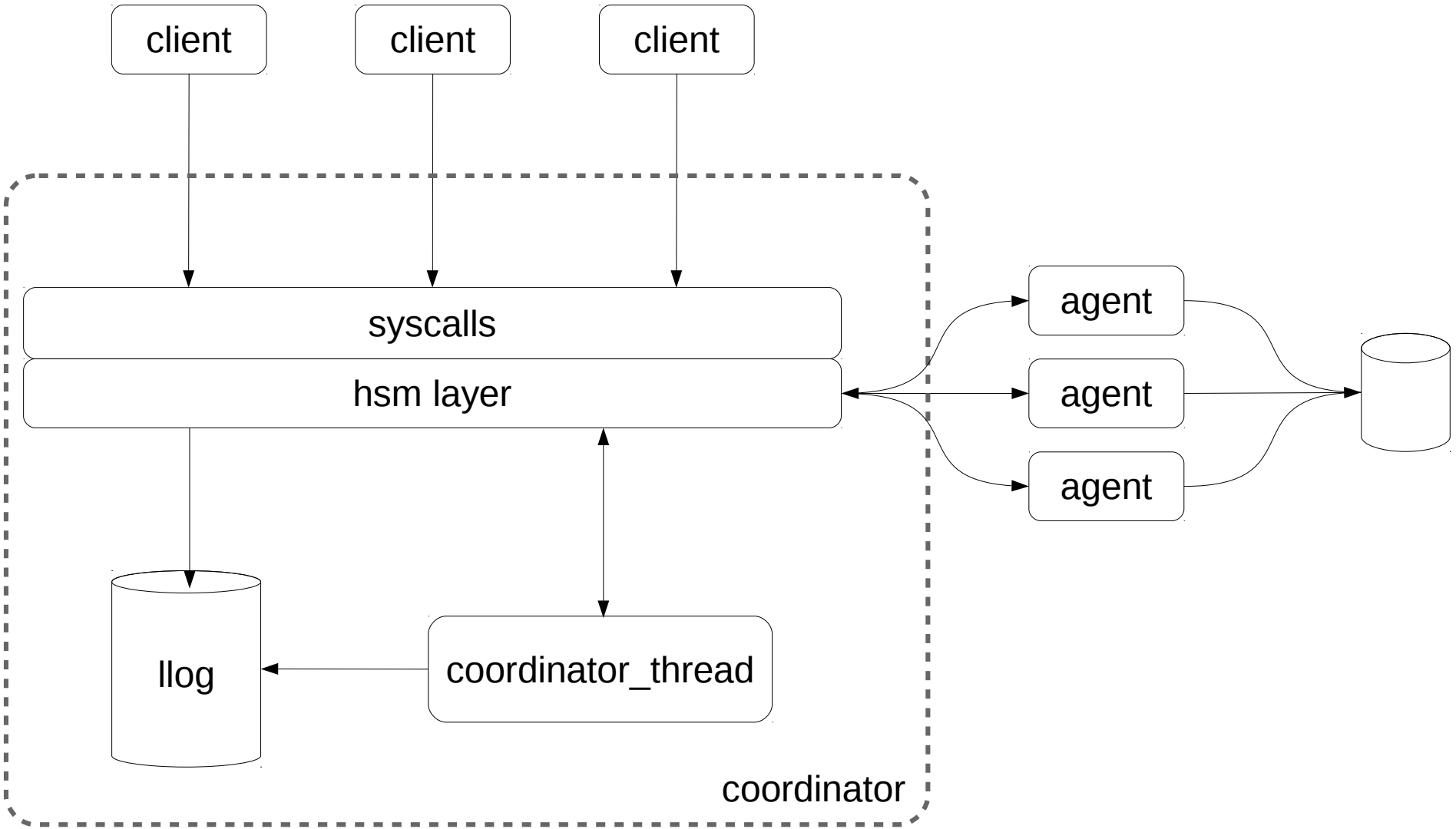
Performance improvements



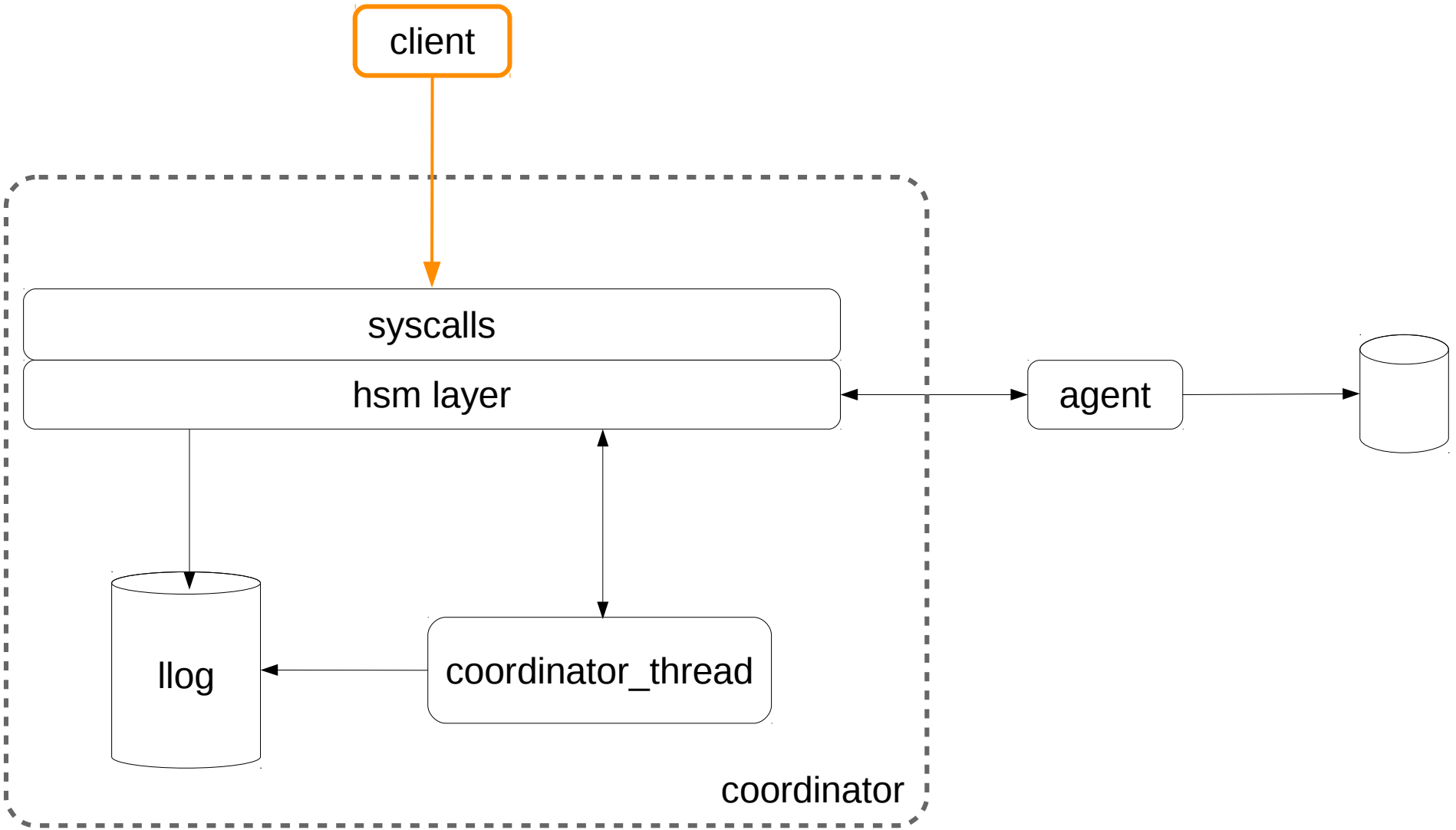
Remaining issues

- Still not scalable (limited at 500 req/s)
- Current behaviour does not fit every workload
- Bugs: LU-5216, LU-5896

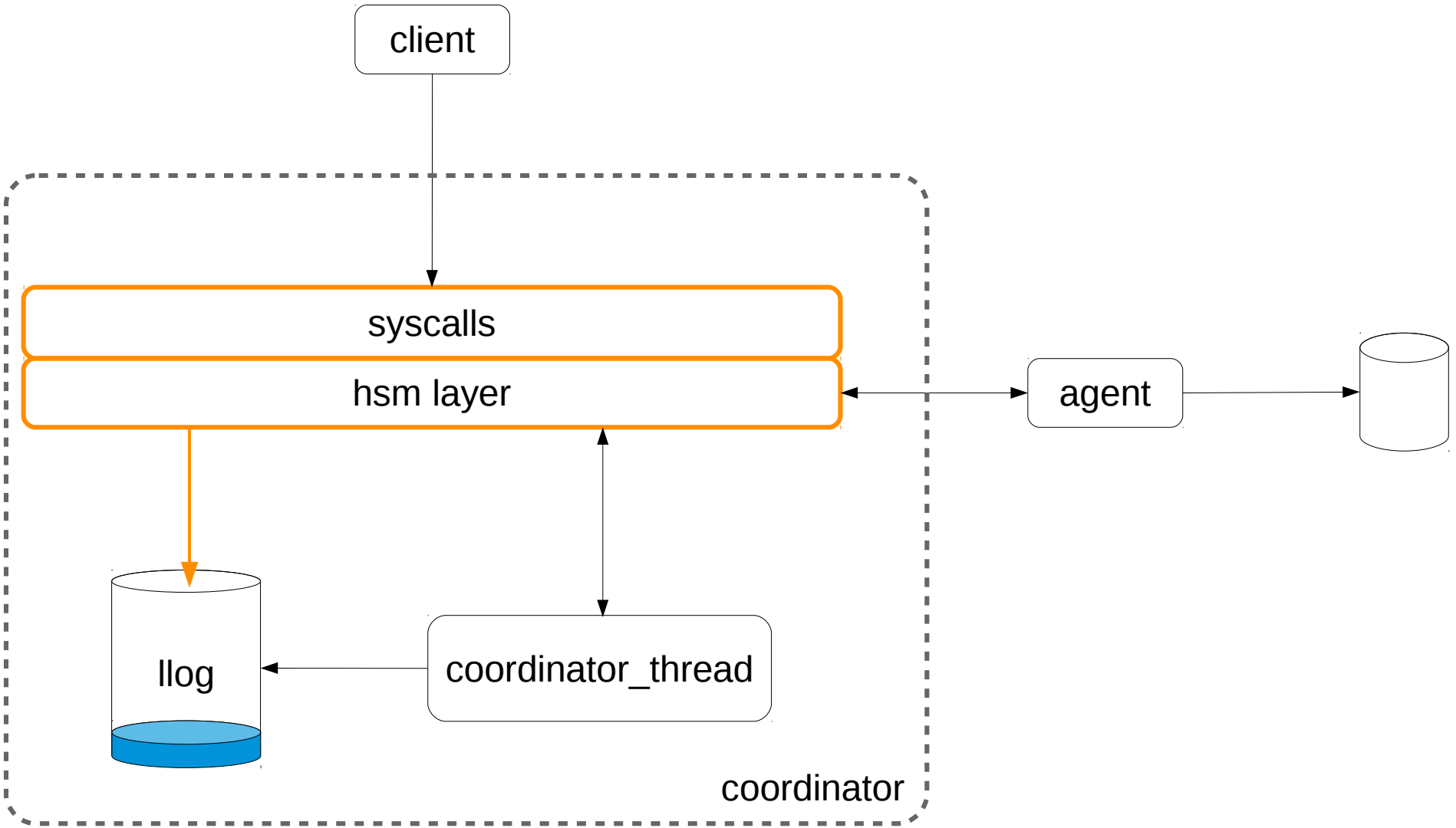
Current design



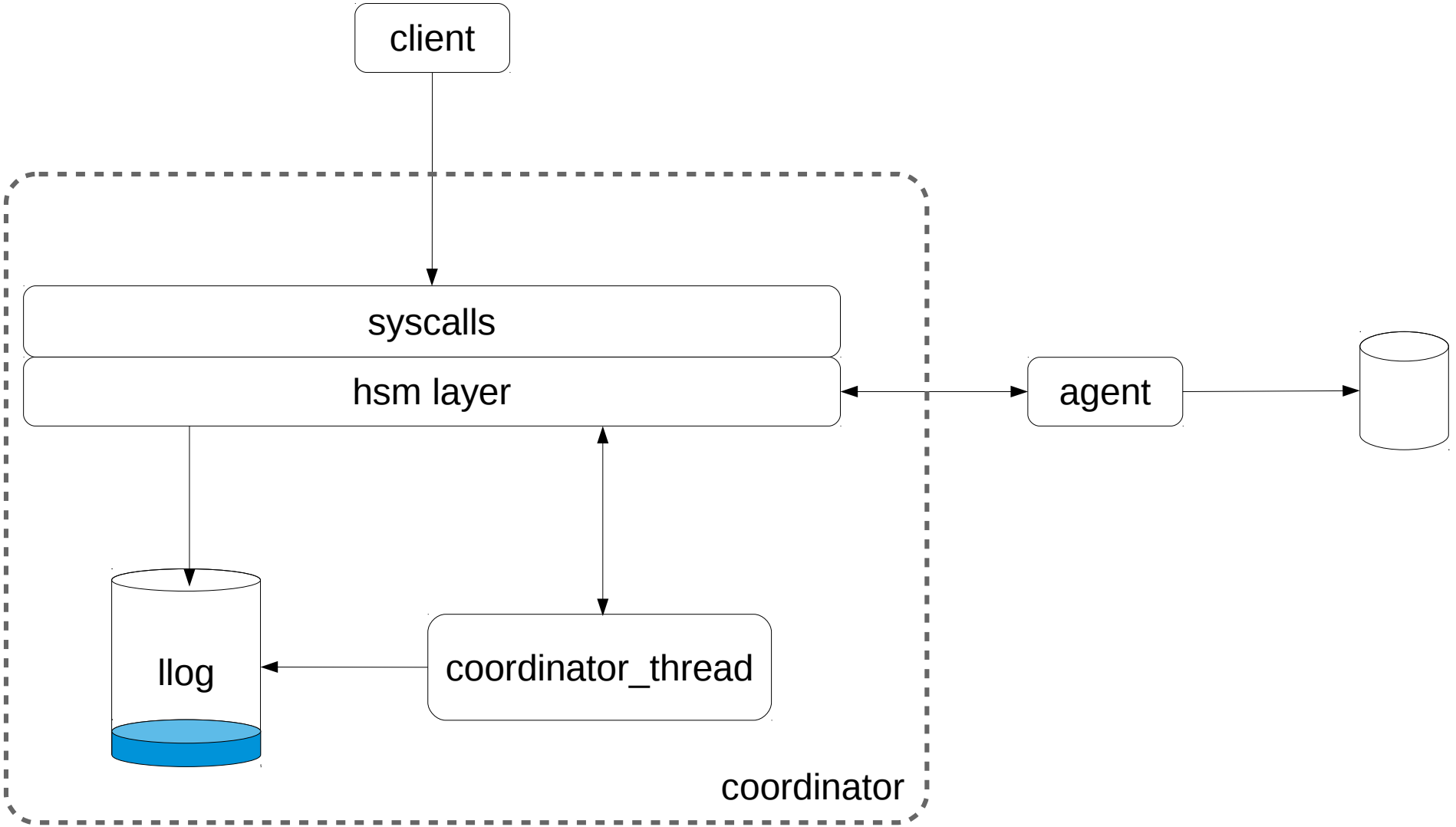
Current design



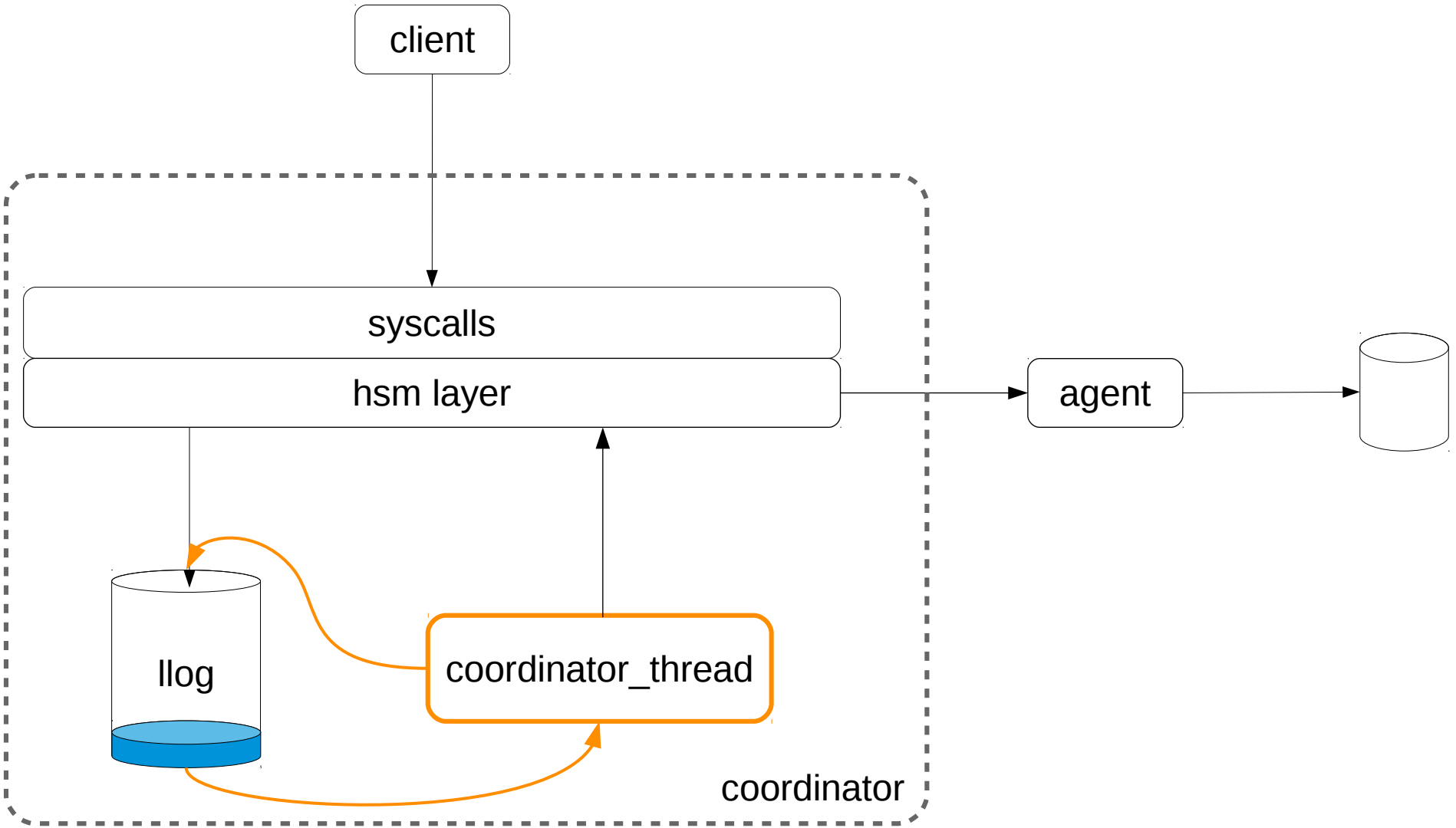
Current design



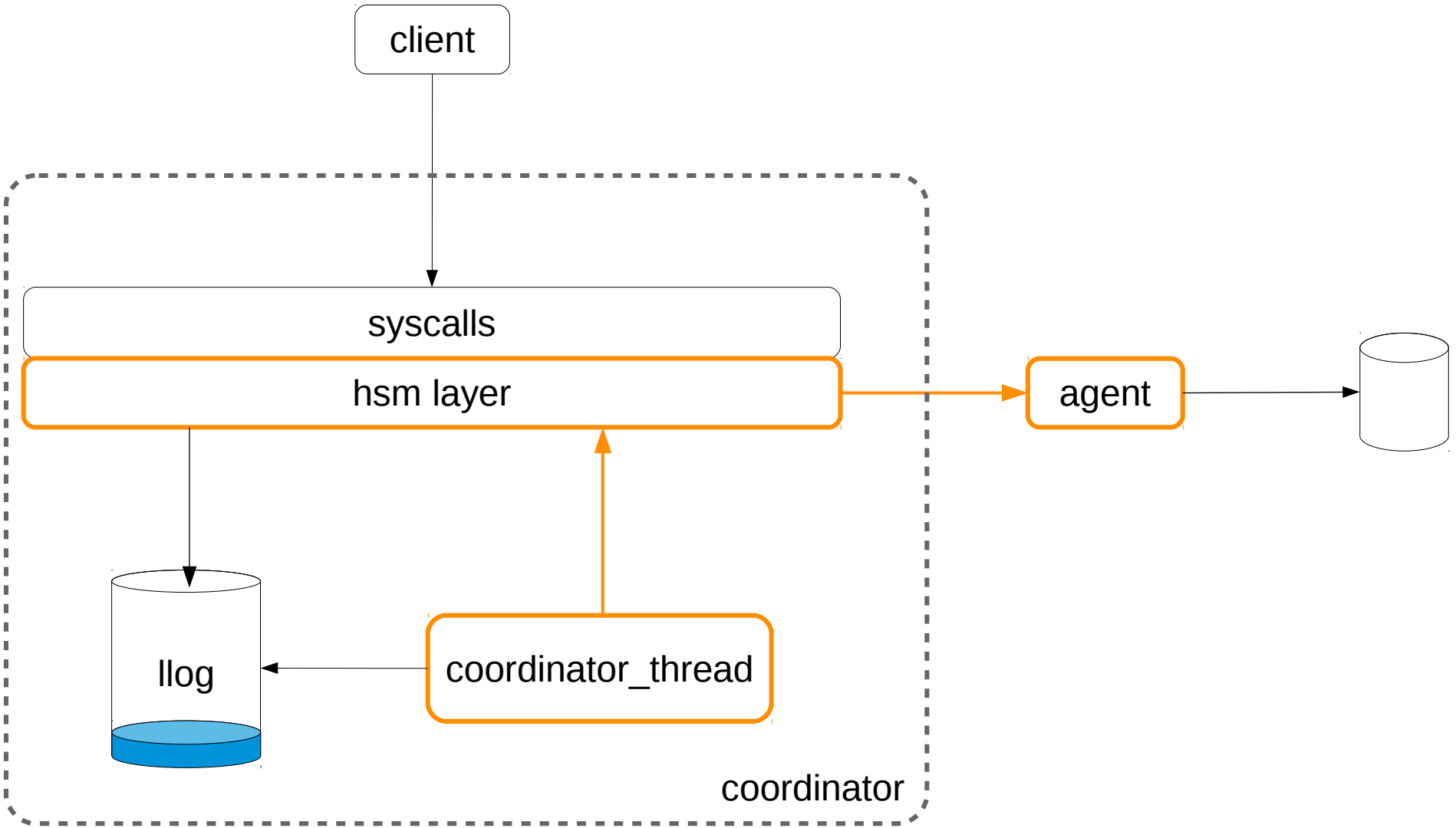
Current design



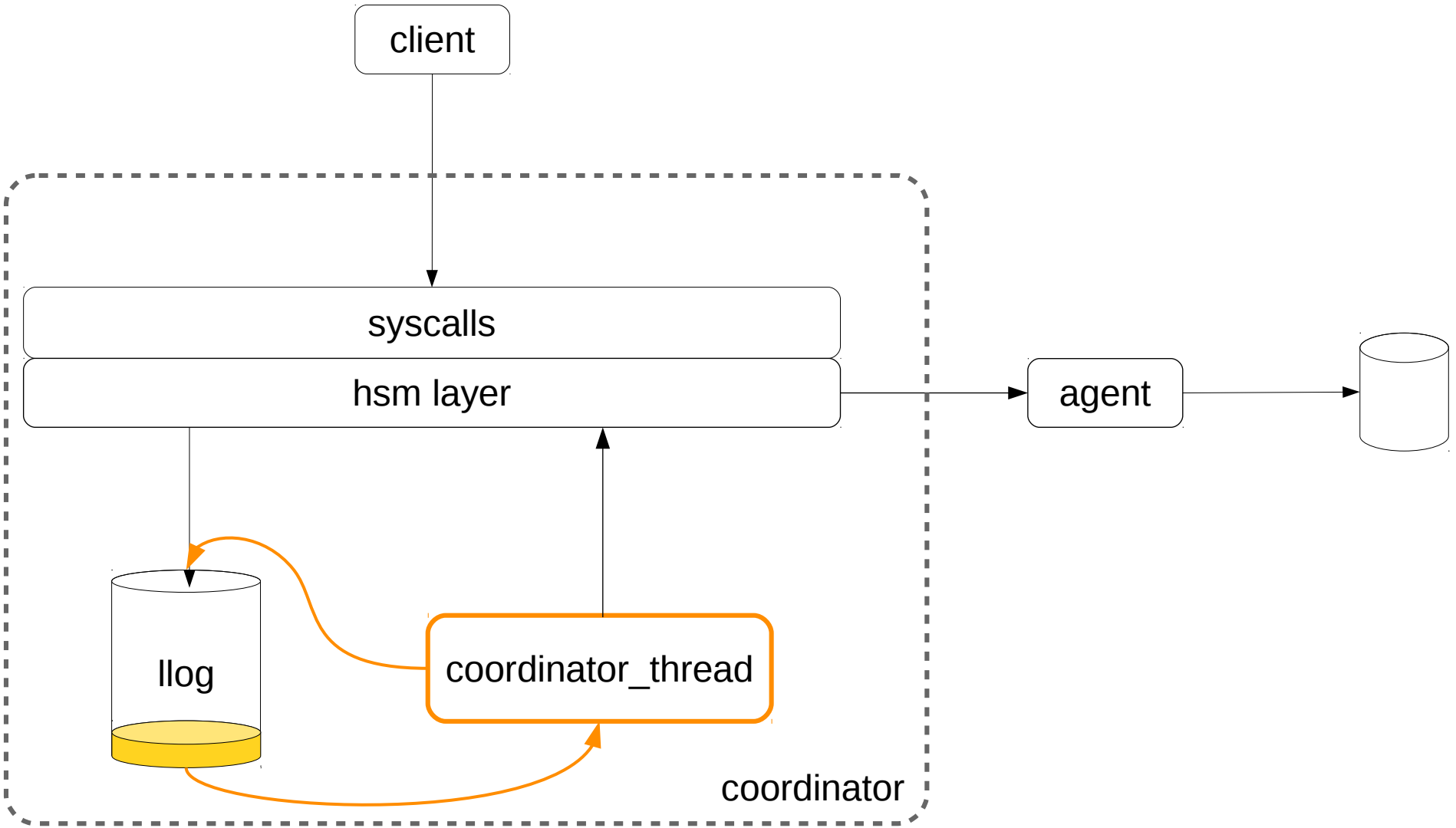
Current design



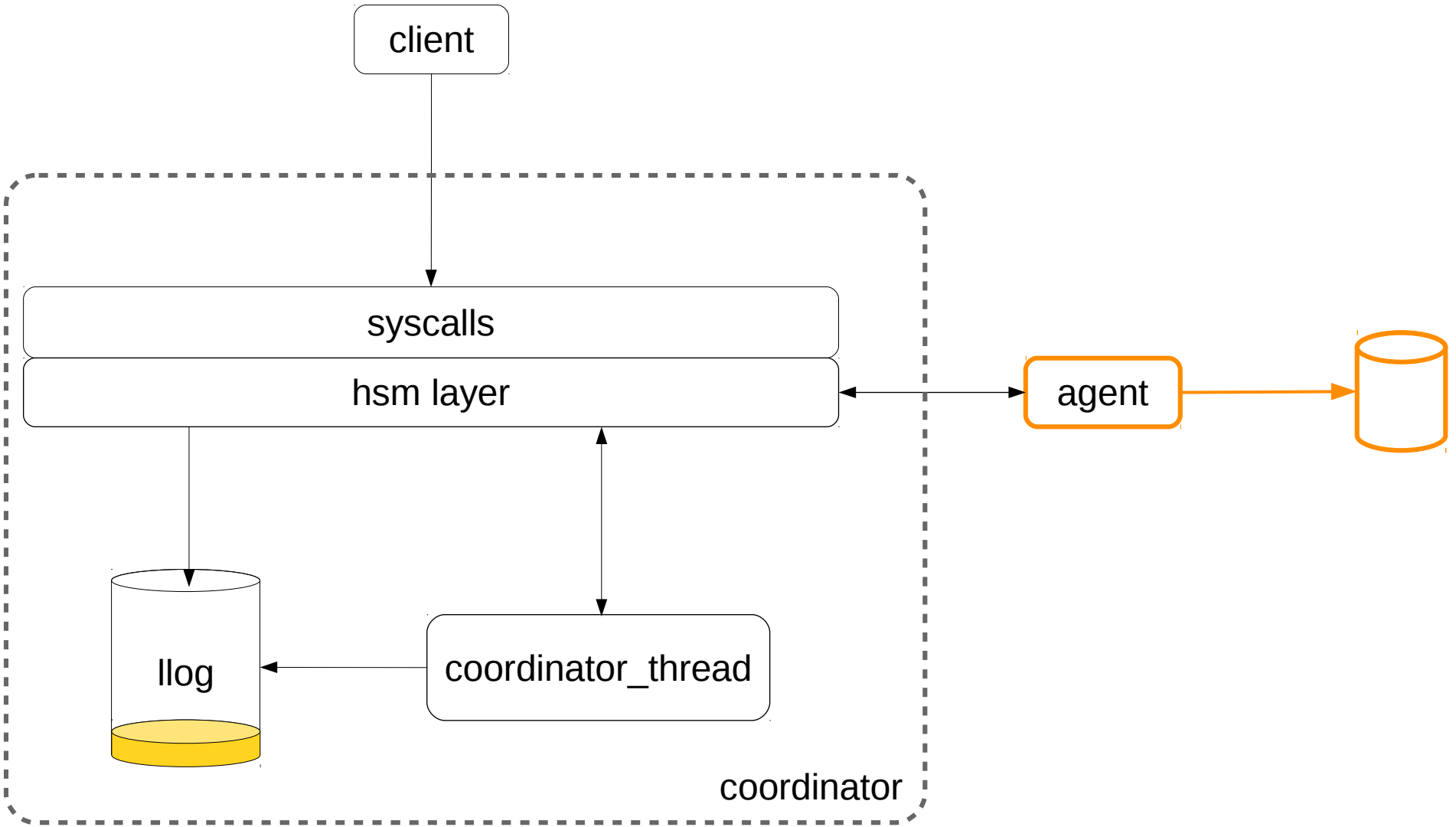
Current design



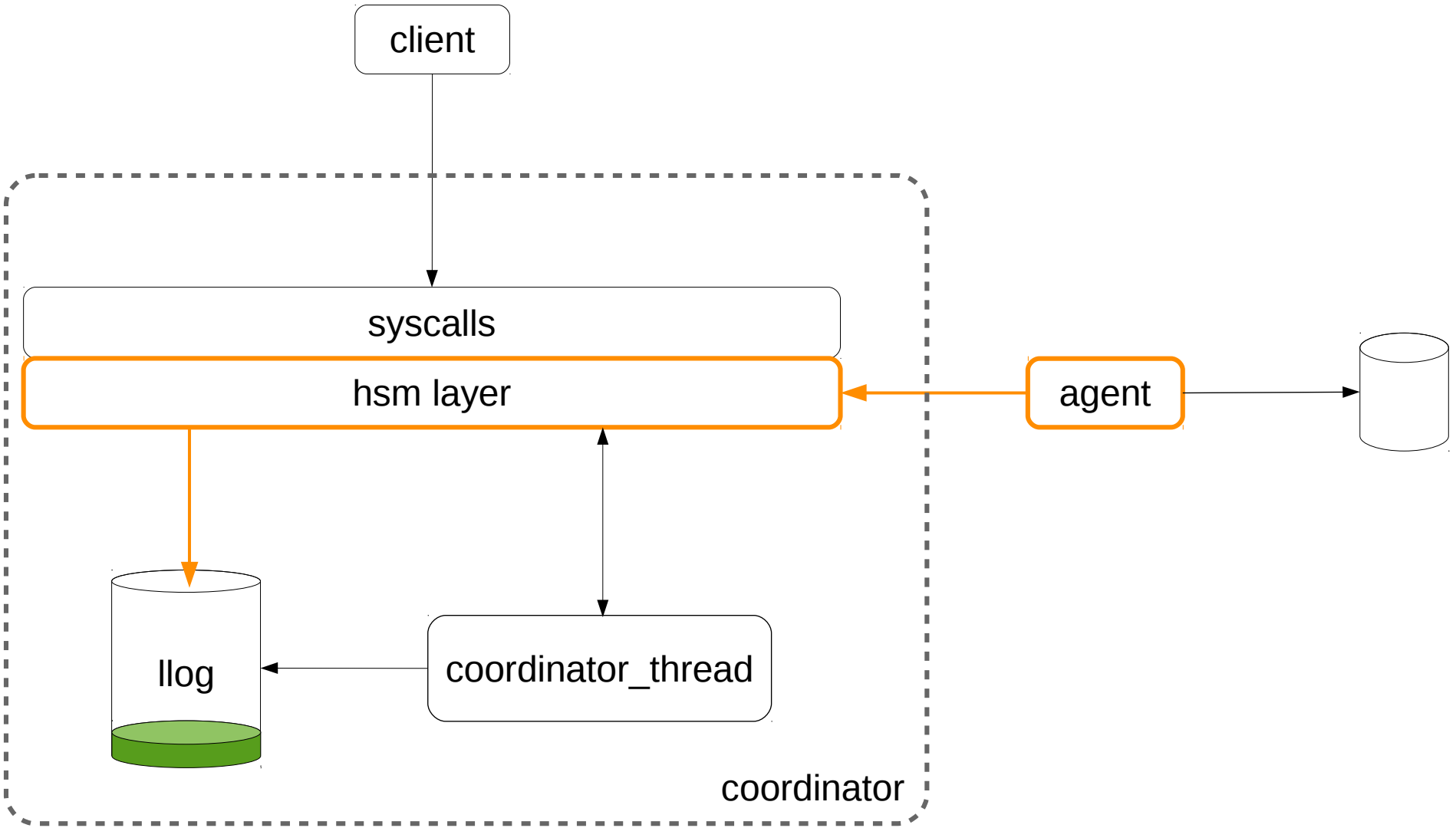
Current design



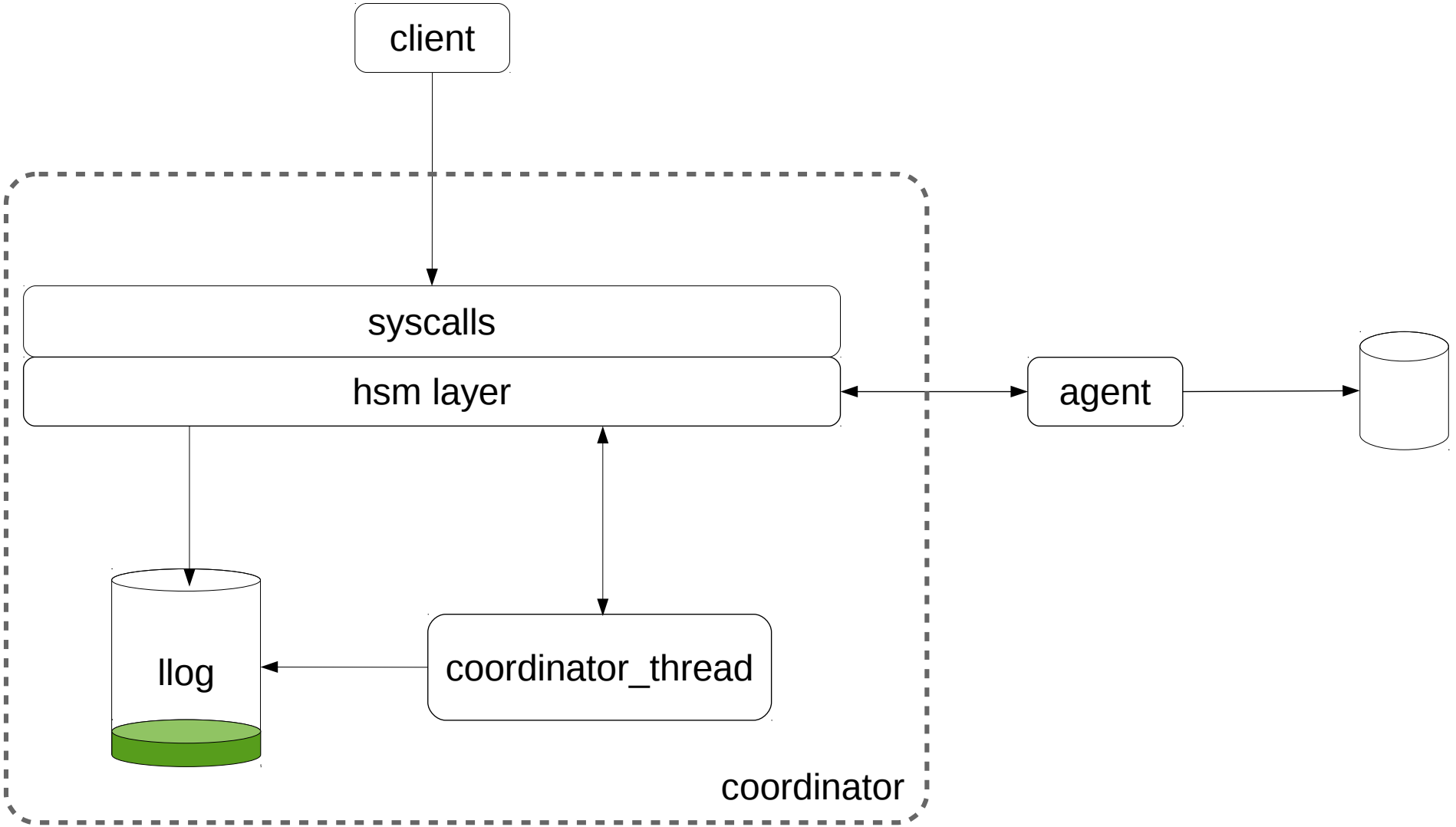
Current design



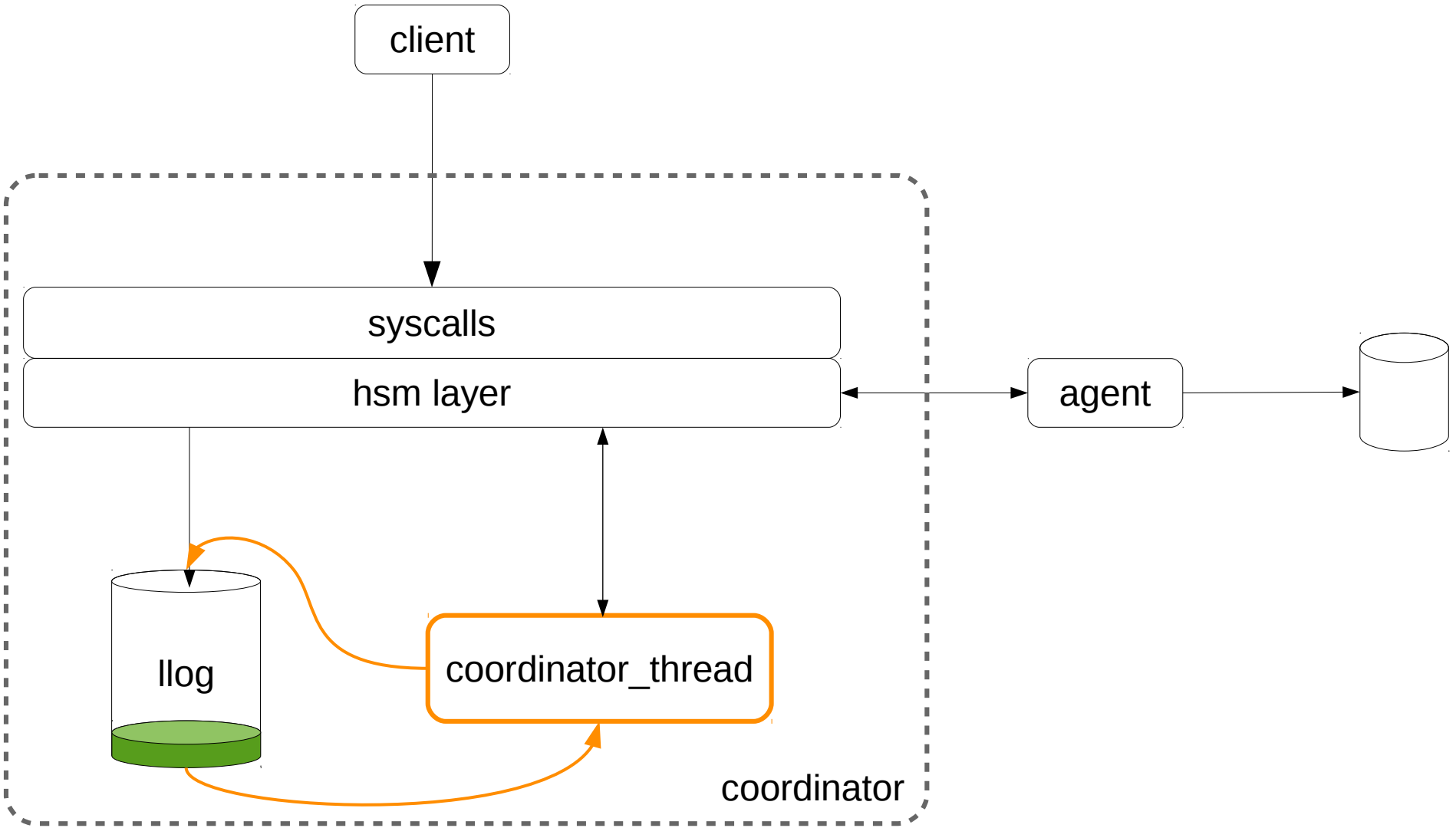
Current design



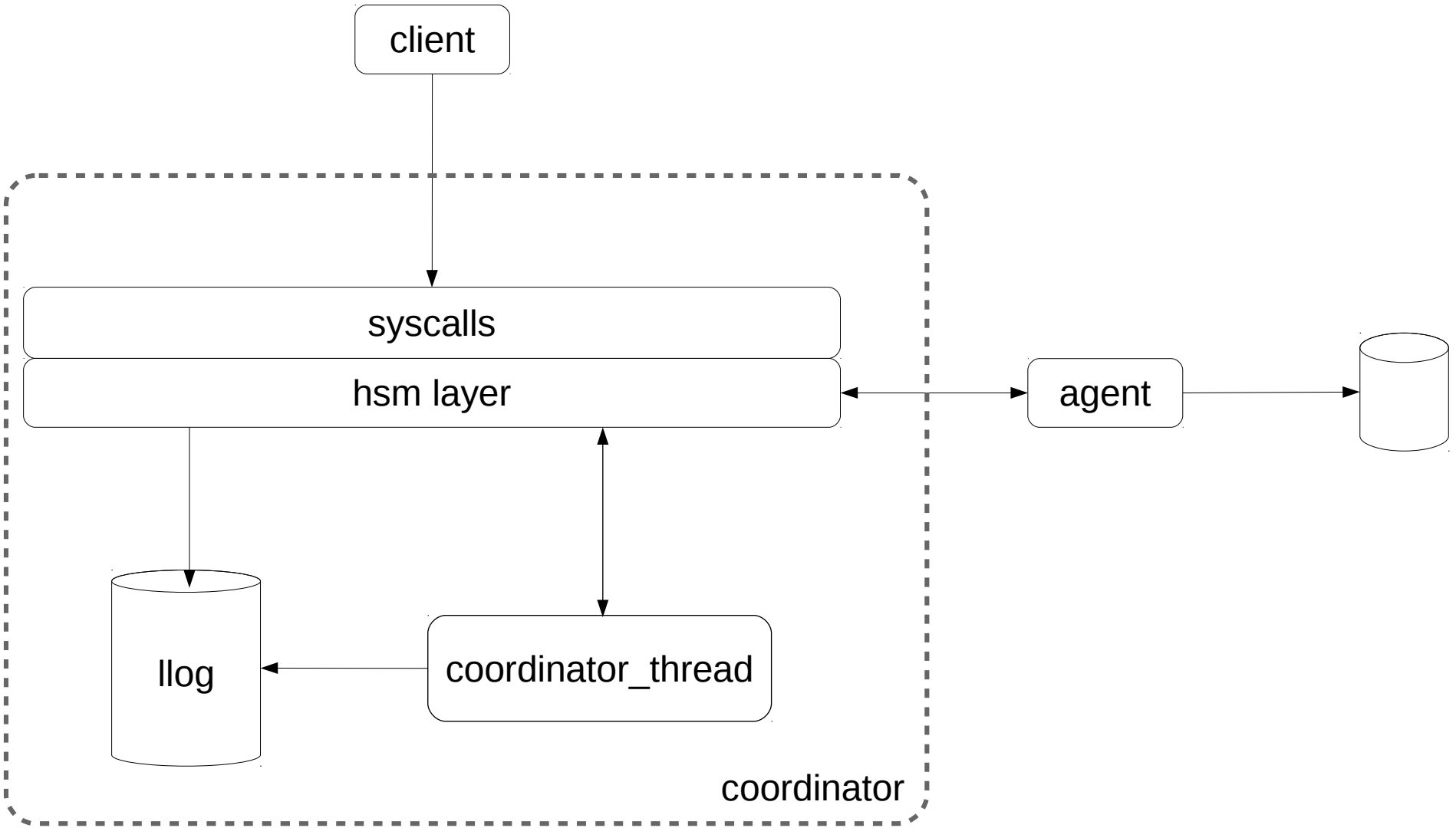
Current design



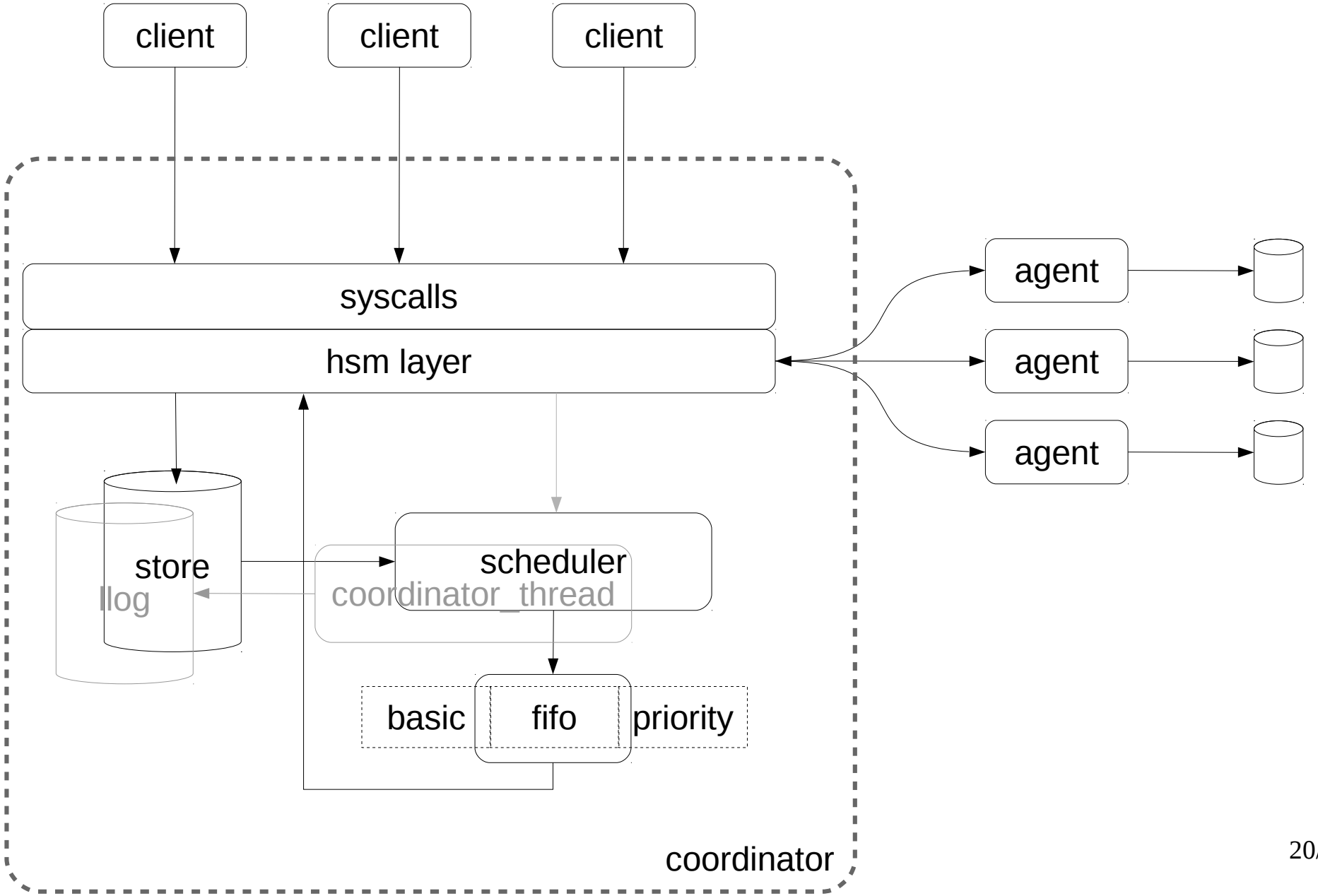
Current design



Current design



Proposal


client
client
client
syscalls
hsm layer
agent
agent
agent
store
llog
scheduler
coordinator_thread
basic
fifo
priority
coordinator

New structures

scheduler:

- declare_request
- cancel_request
- register_agent
- unregister_agent

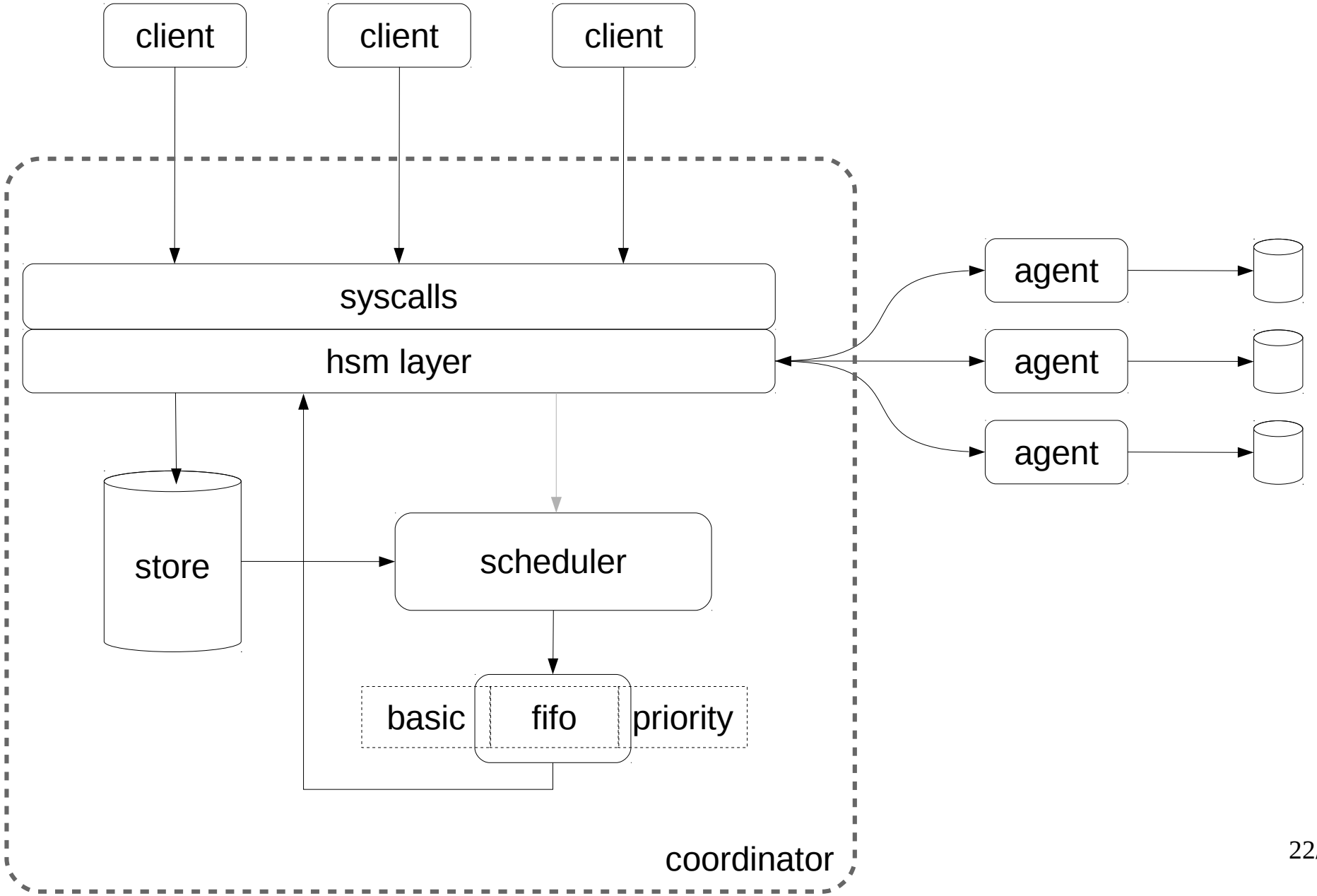
store:

- declare_request
- cancel_request

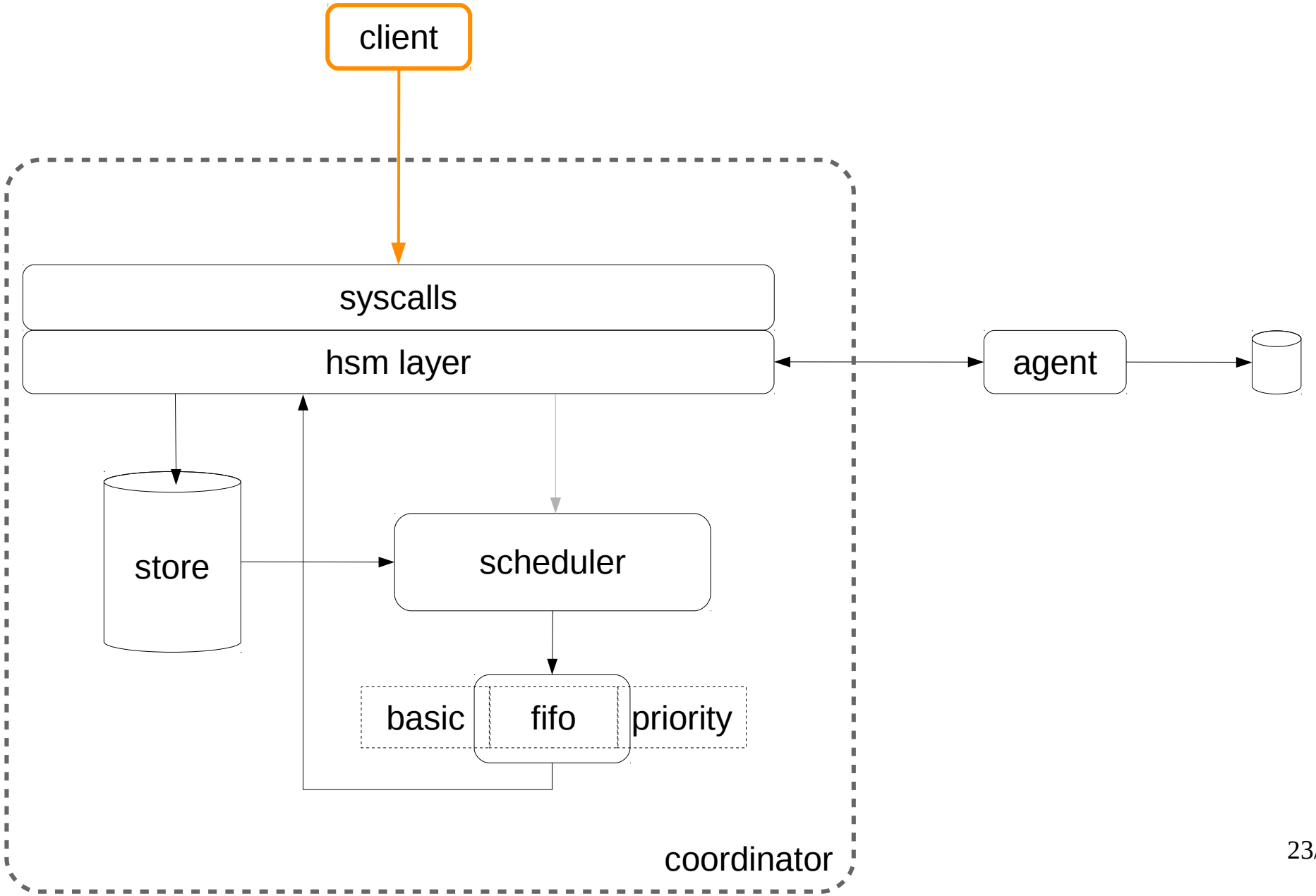
iterator:

- next

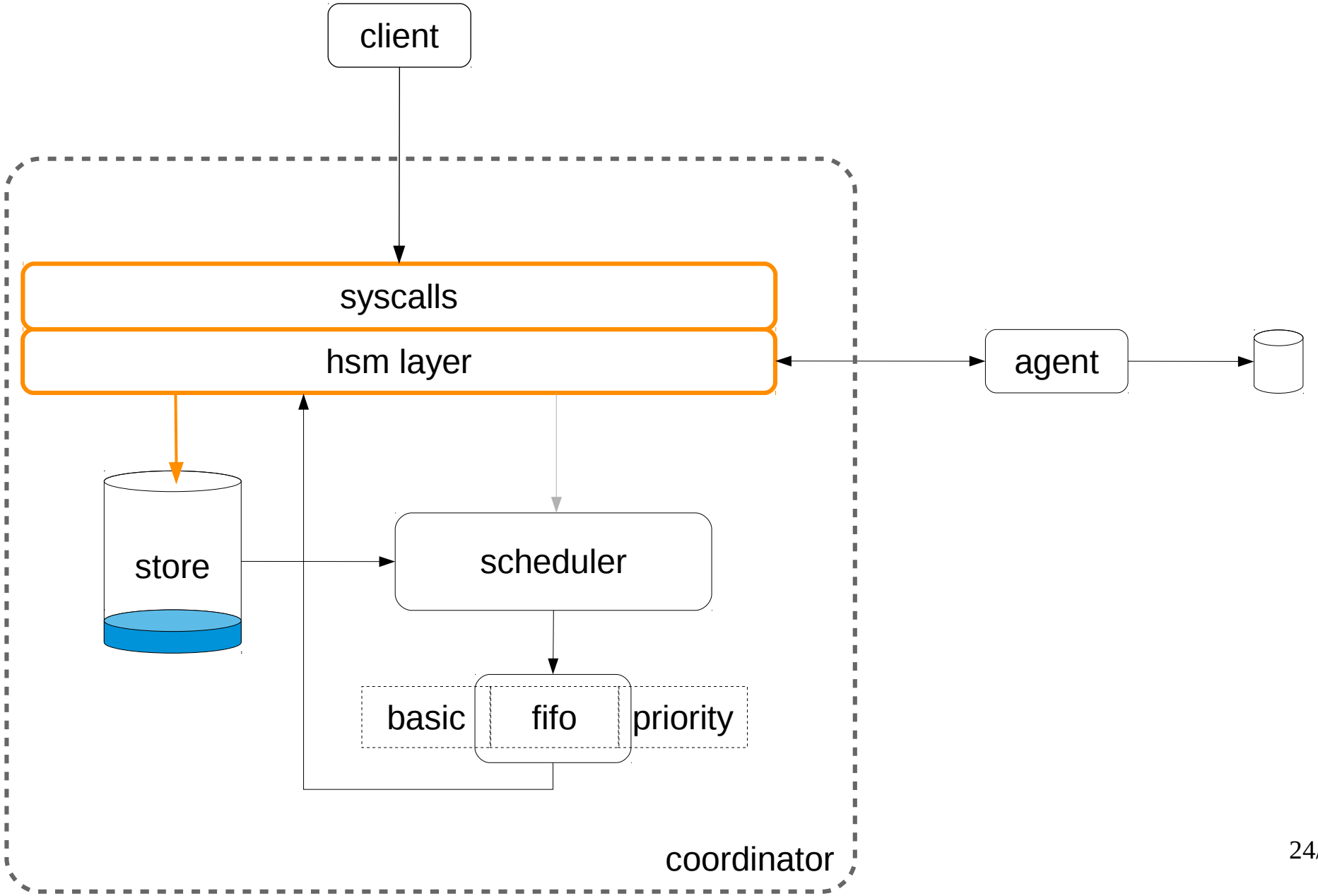
Proposal


client
client
client
syscalls
hsm layer
agent
agent
agent
store
scheduler
basic
fifo
priority
coordinator

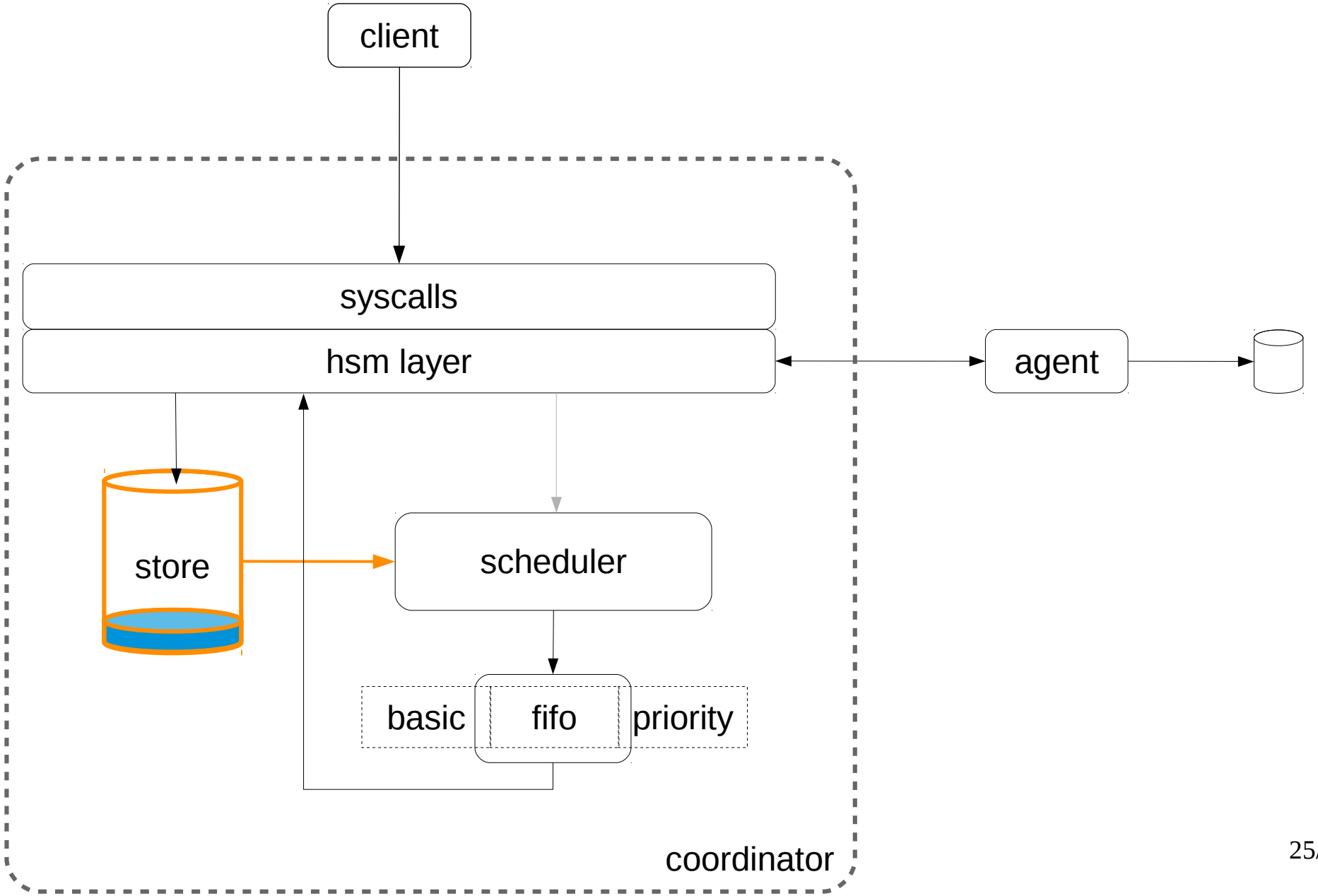
Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

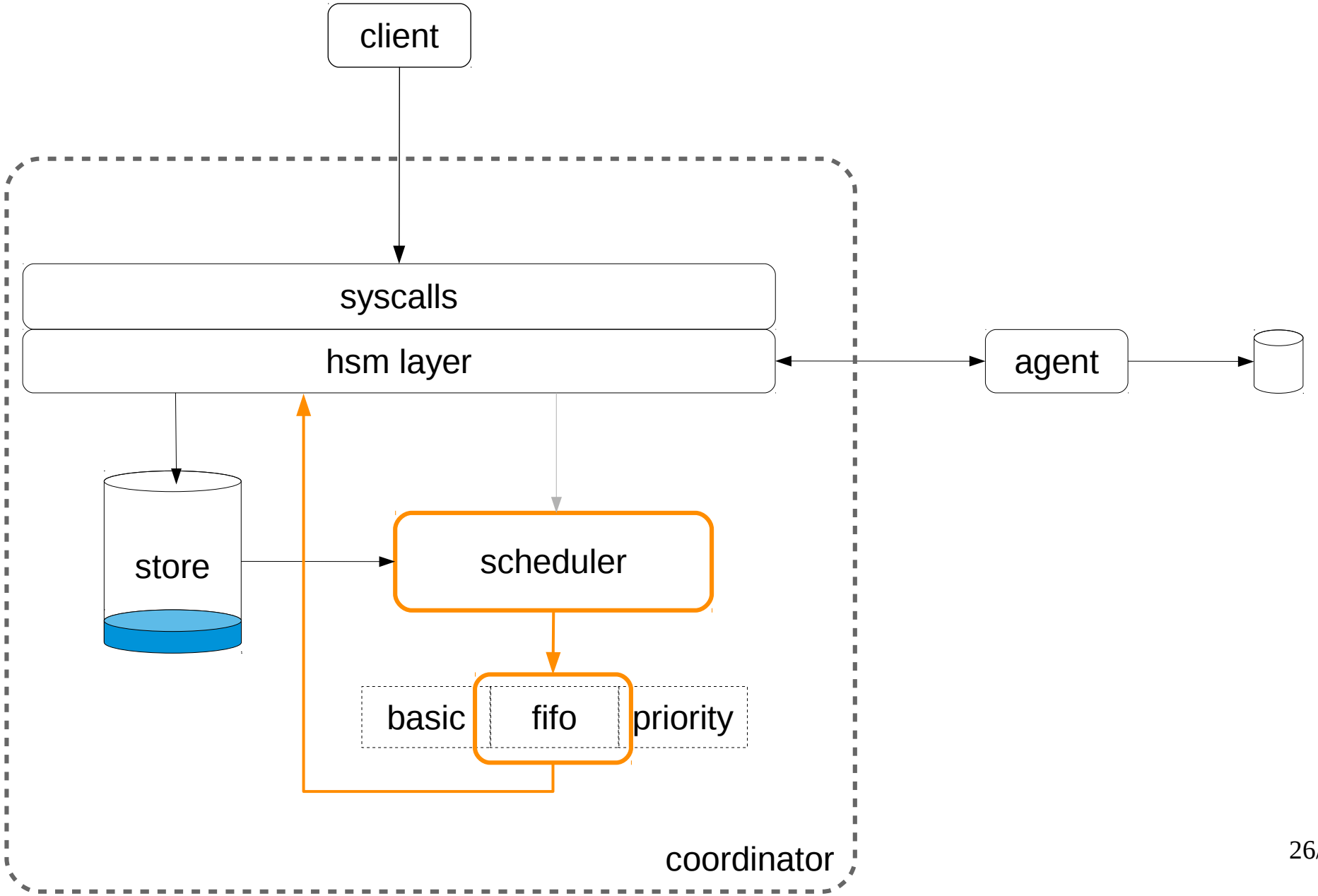
Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

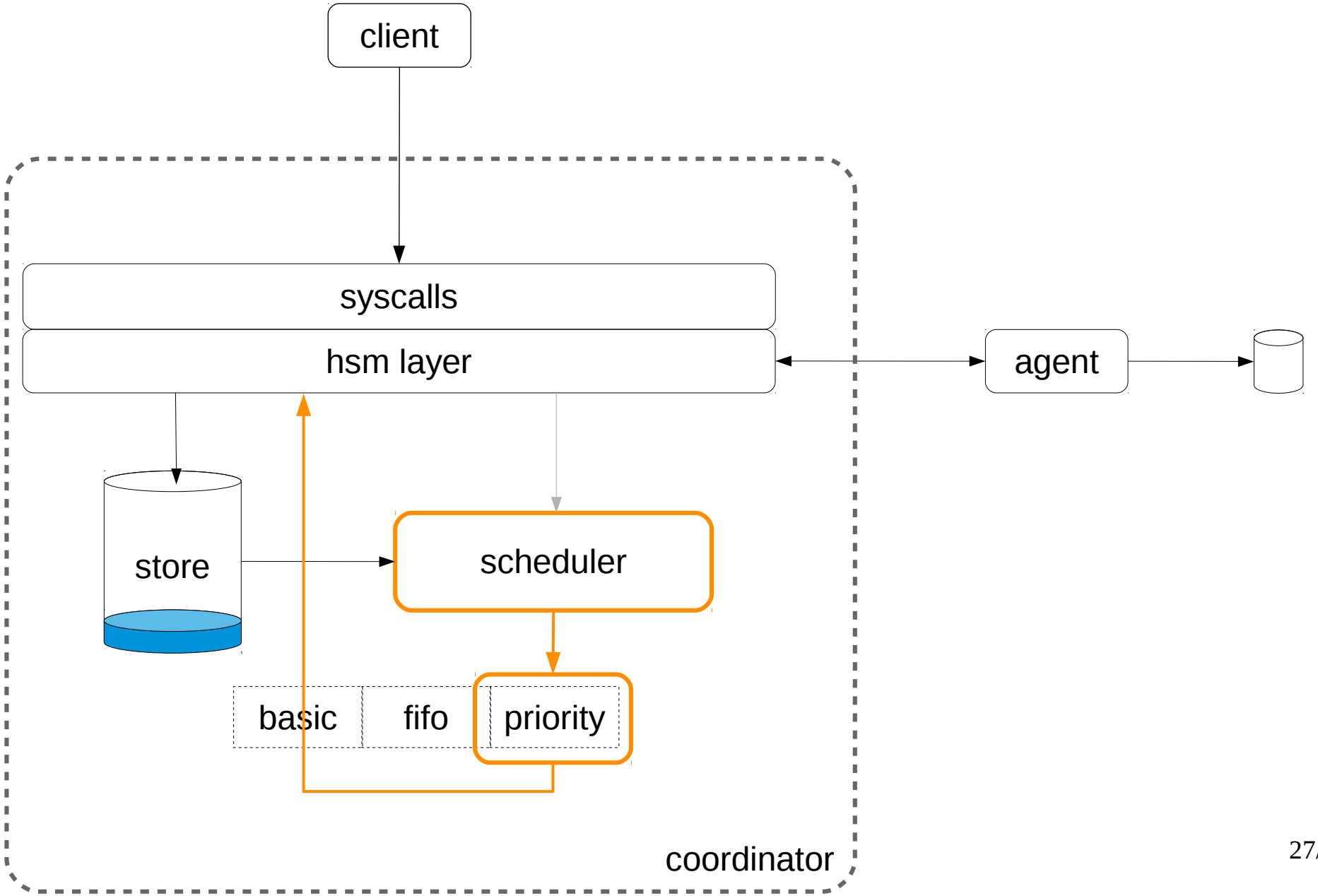
Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

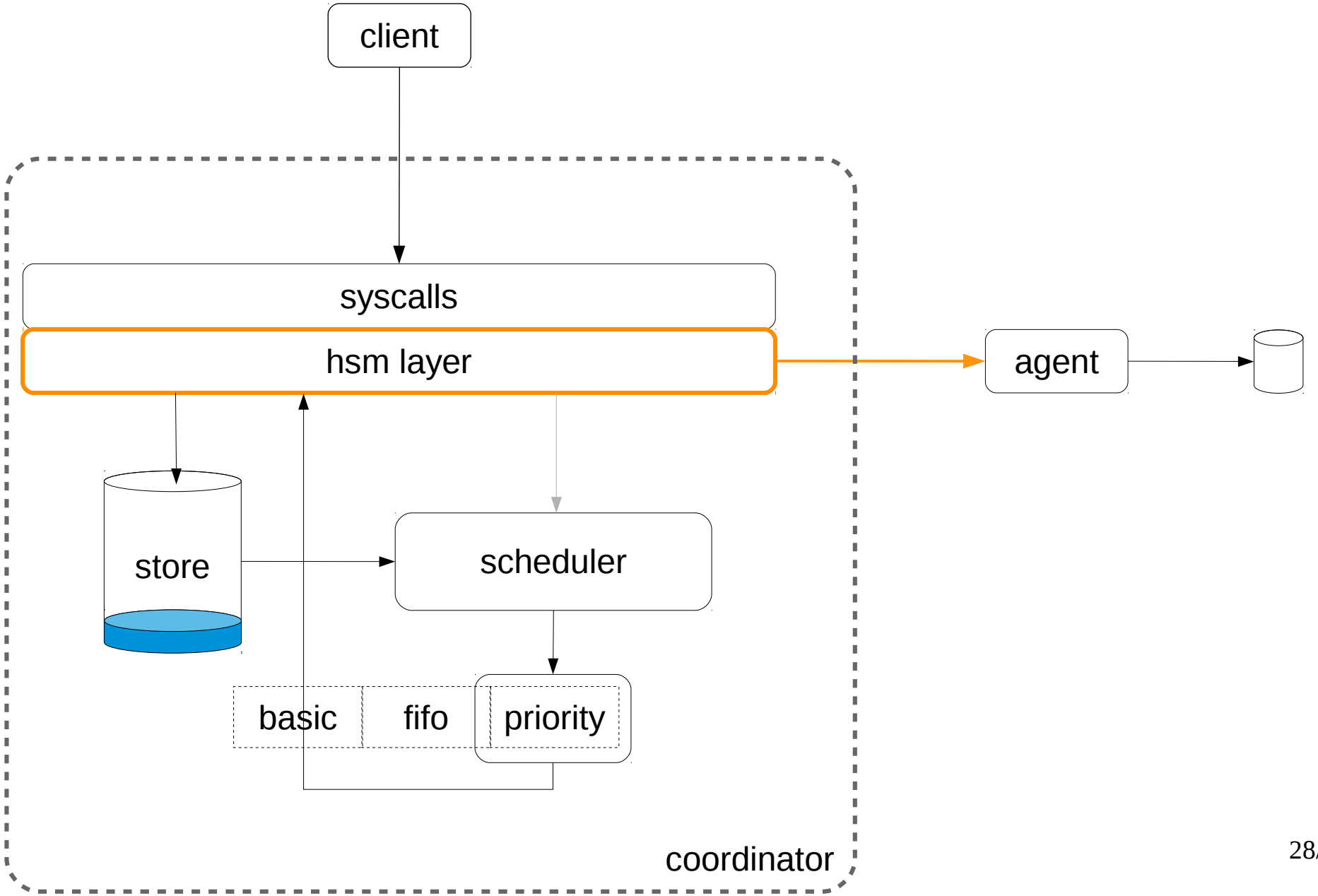
Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

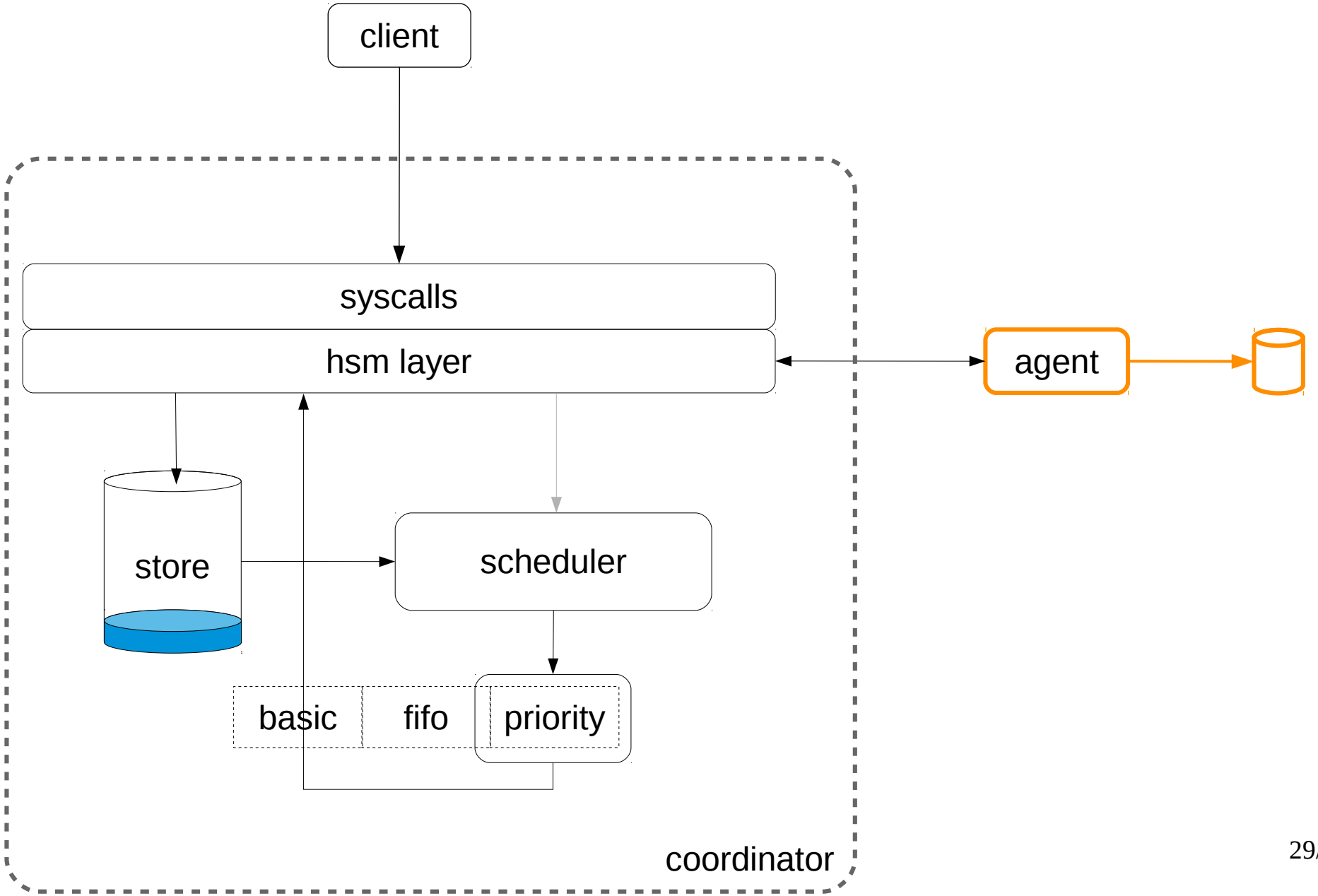
Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

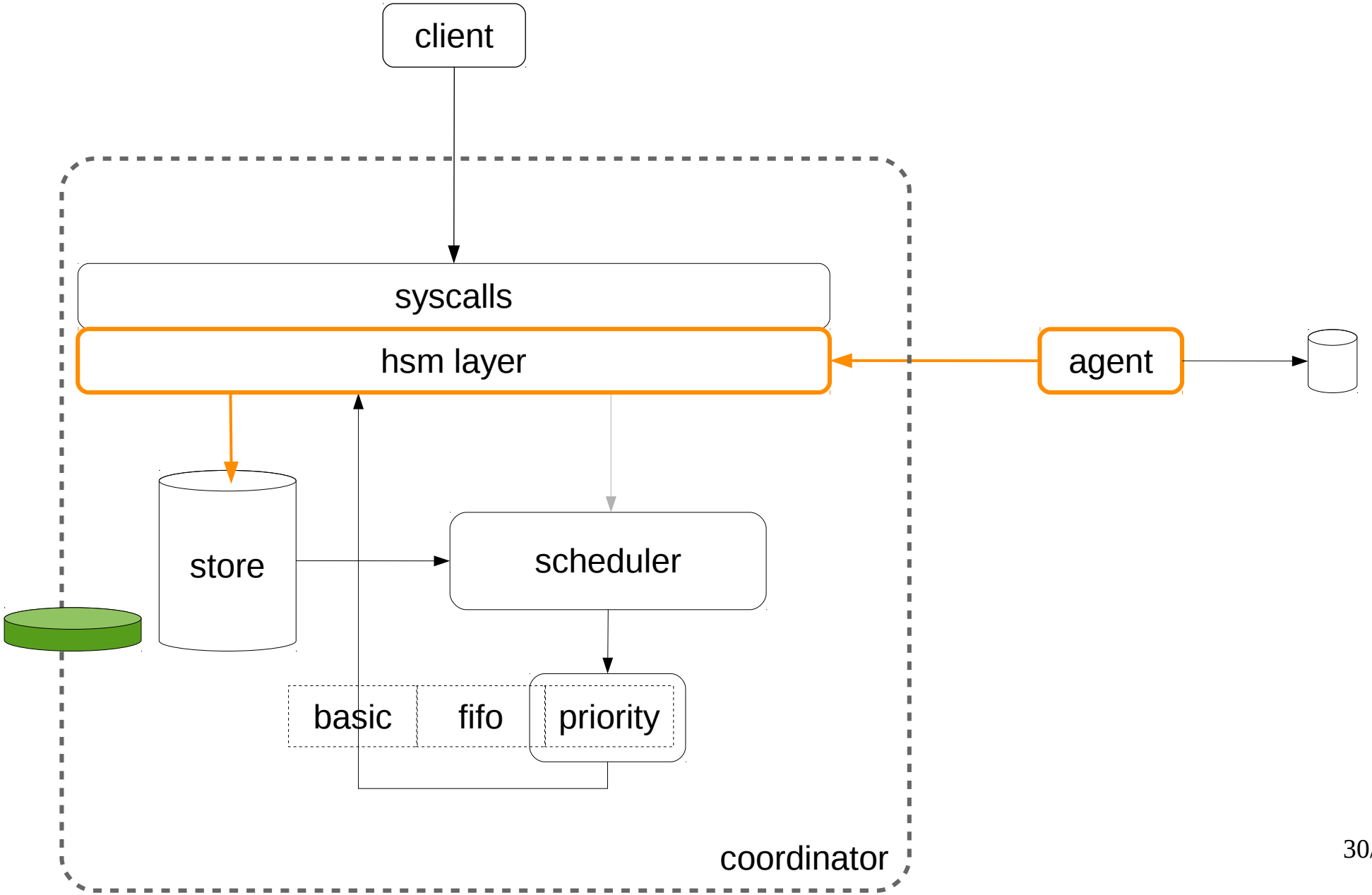
Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

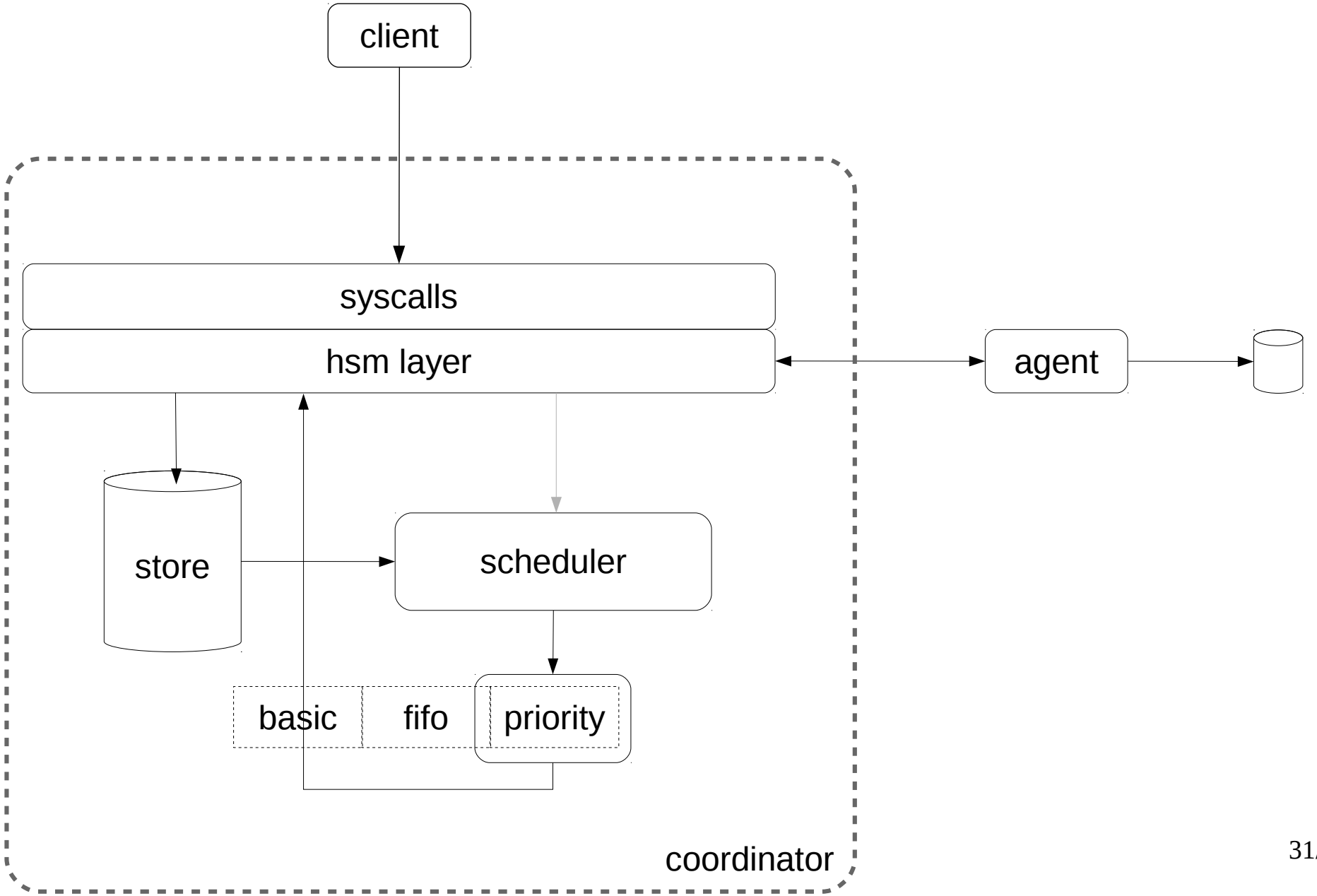
Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

Proposal


client
syscalls
hsm layer
agent
store
scheduler
basic
fifo
priority
coordinator

Pros and cons

+ scalable
+ tunable / improvable
+ cleaner code
+ new features

– hard to integrate
– lots of coding / review

DE LA RECHERCHE À L'INDUSTRIE

cea

Thank you!

Integration planning

1. Clean split between hsm and mdt
2. General code cleanup
3. Remove completed requests from the llog
4. Stop storing the requests' state in the llog
5. Replace the llog (?)
6. Move on to loadable policies