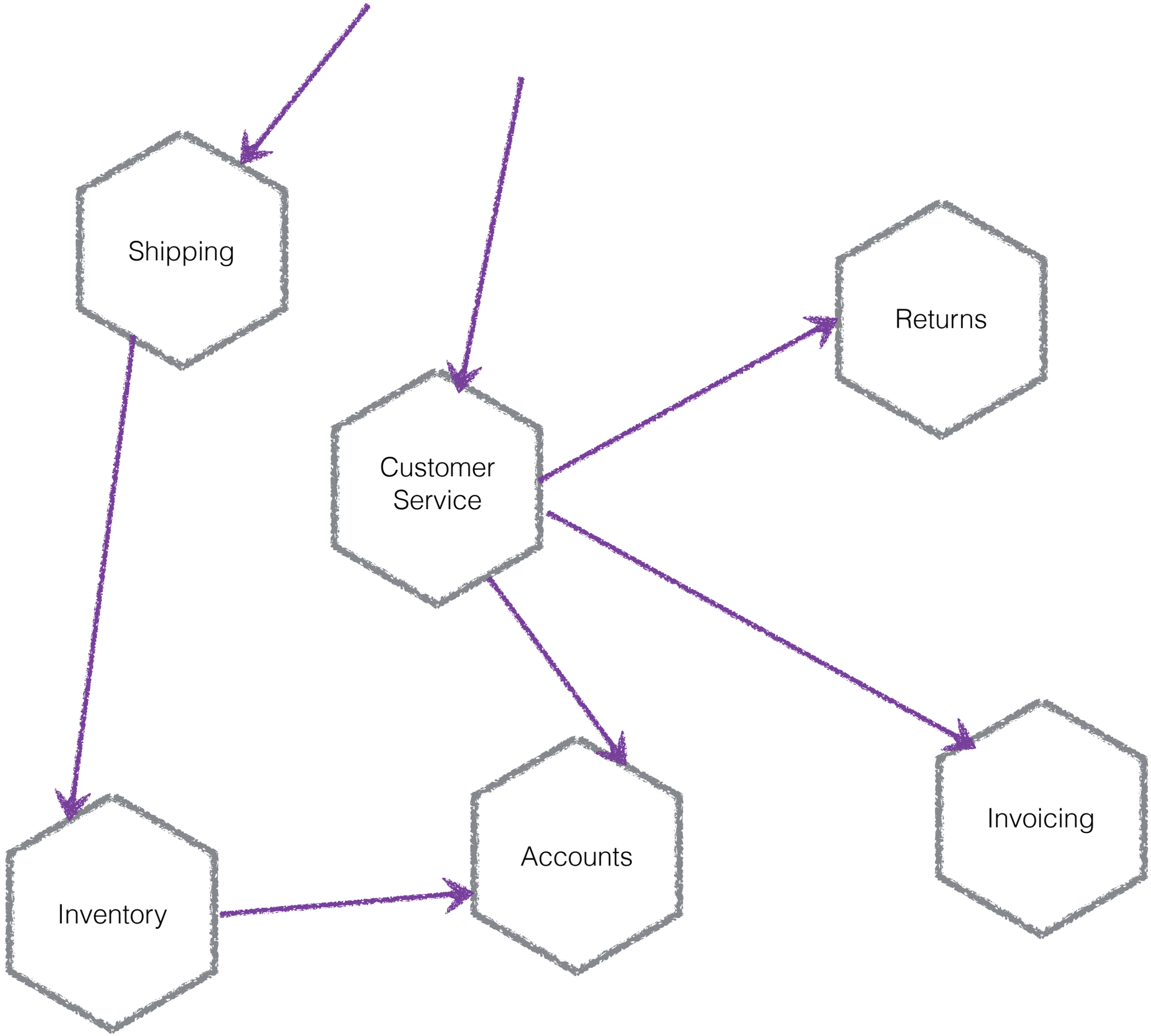
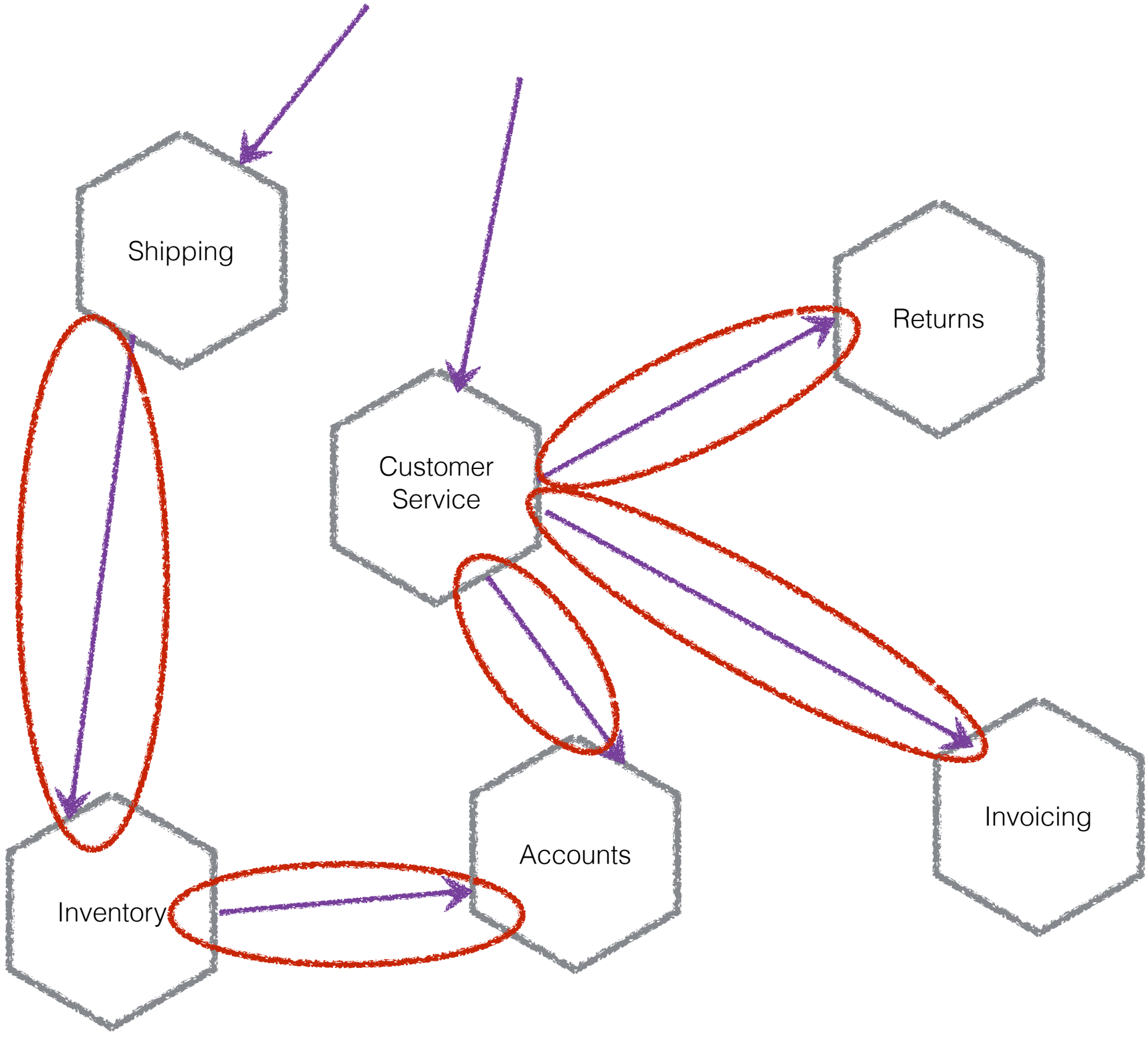
MONOLITHS

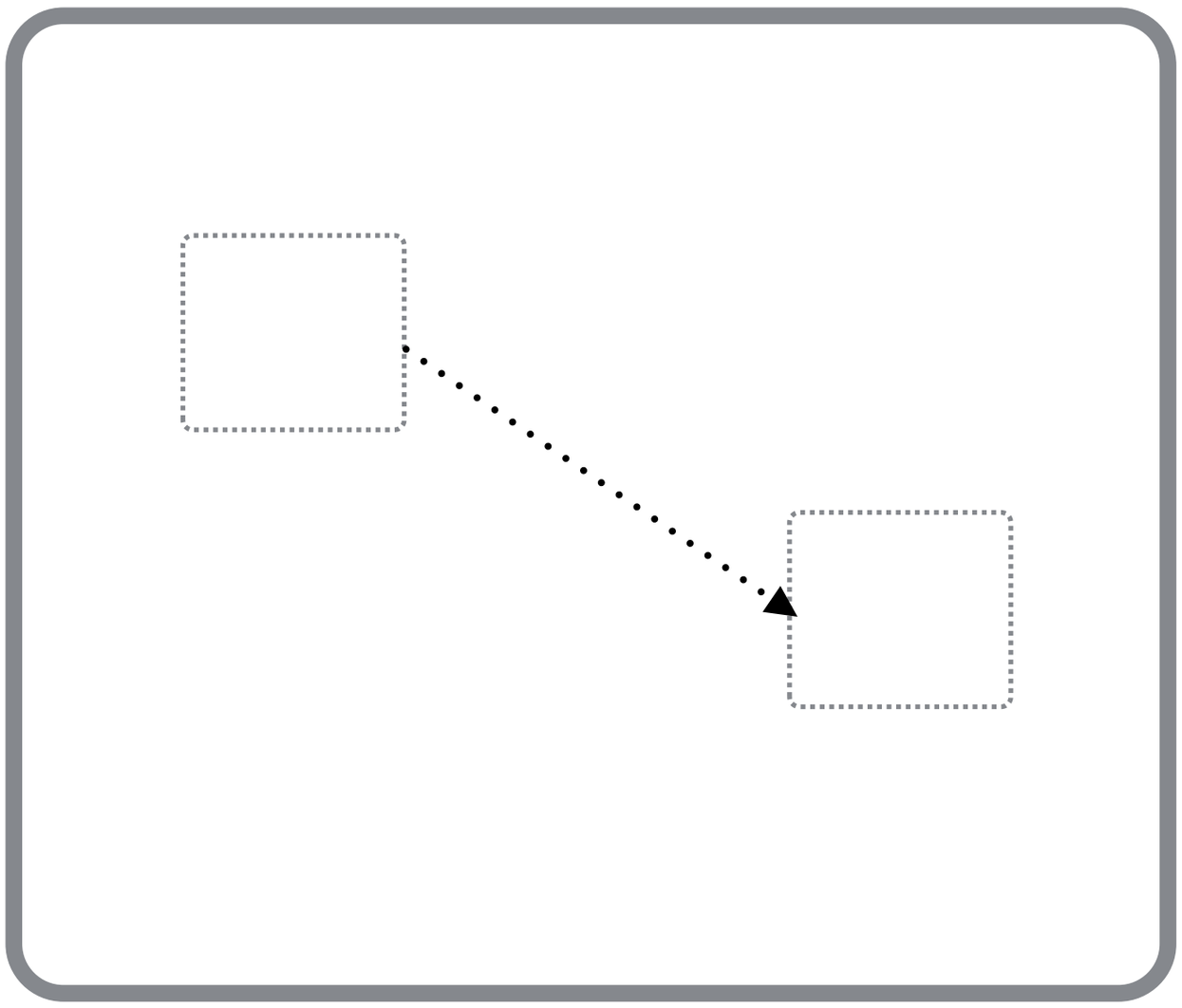
TO MICROSERVICES

Sam Newman

Service Interactions

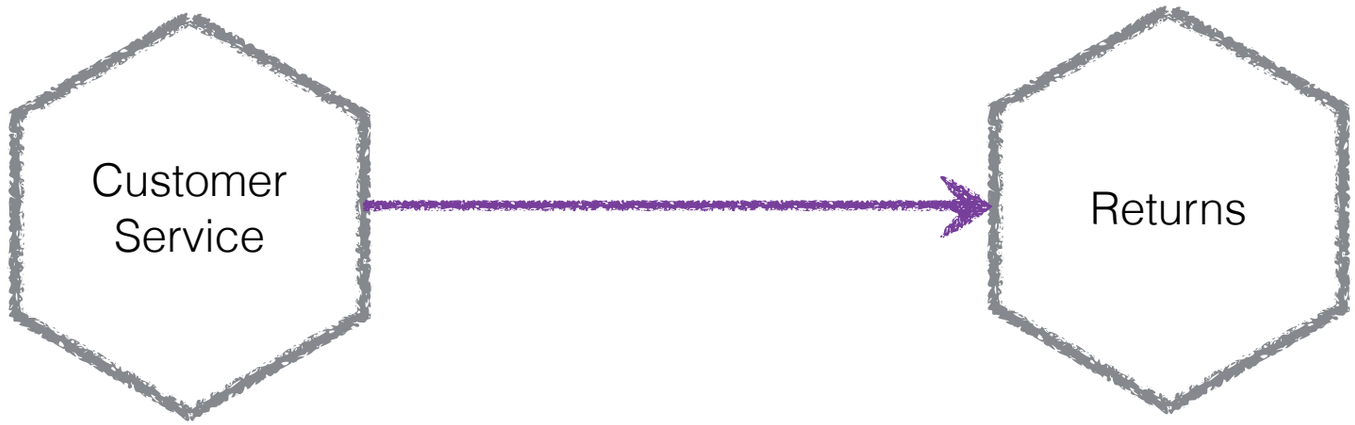

Shipping
Customer Service
Returns
Inventory
Accounts
Invoicing


Shipping
Customer Service
Returns
Inventory
Accounts
Invoicing



Cost of change
is low

Easy to reason
about



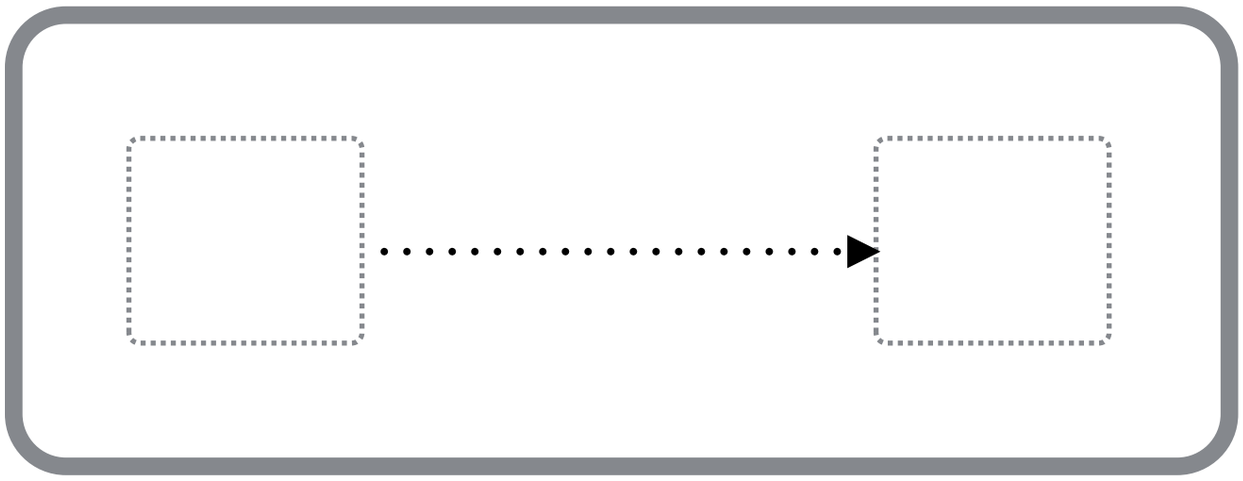
Changing a call ?

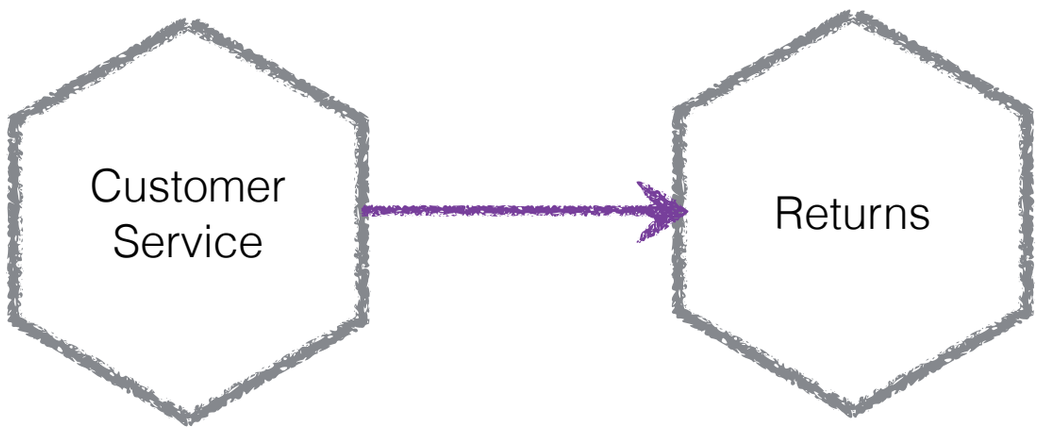
potential API breakage

two deployments to rollout a change

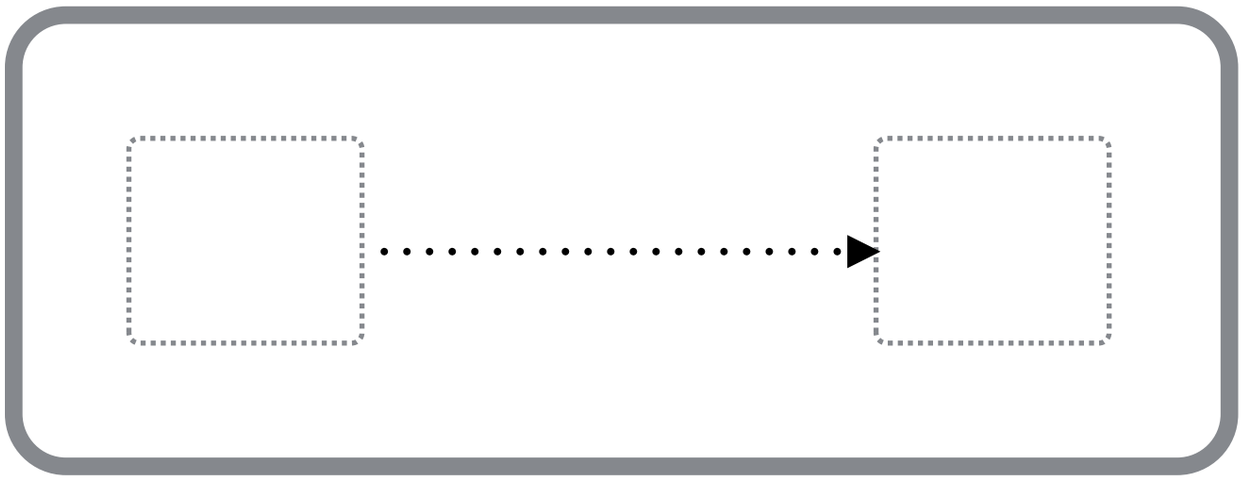
Are calls between services like calls inside a process boundary?

PERFORMANCE IMPLICATIONS

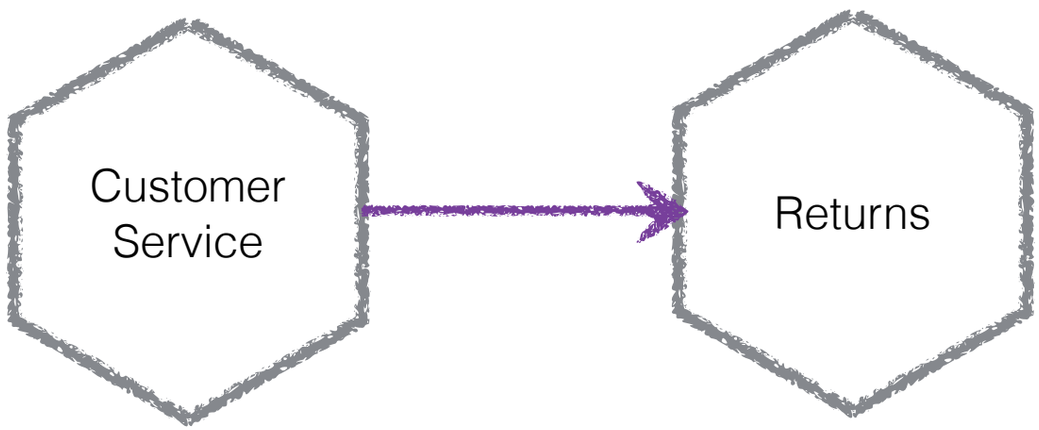




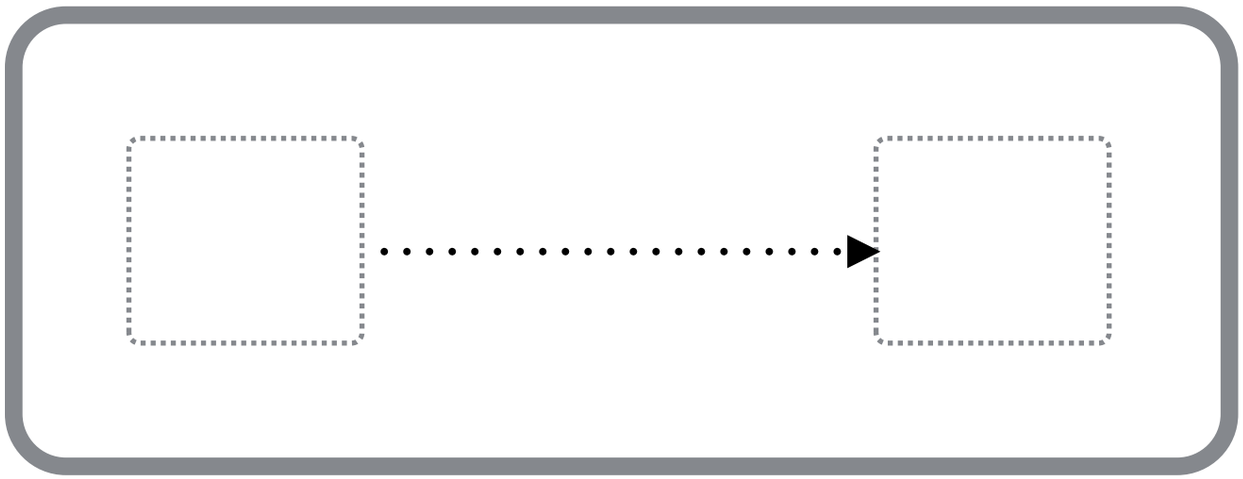
PERFORMANCE IMPLICATIONS



Per-call overhead is very low

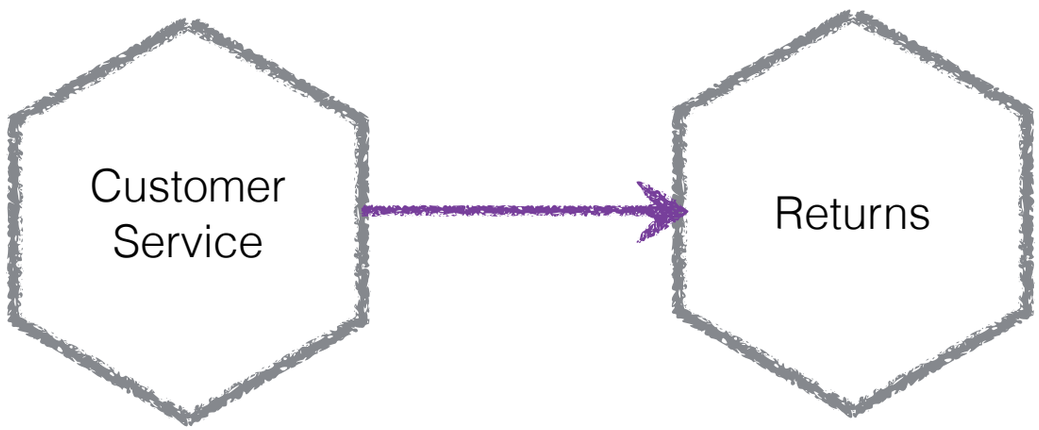


PERFORMANCE IMPLICATIONS

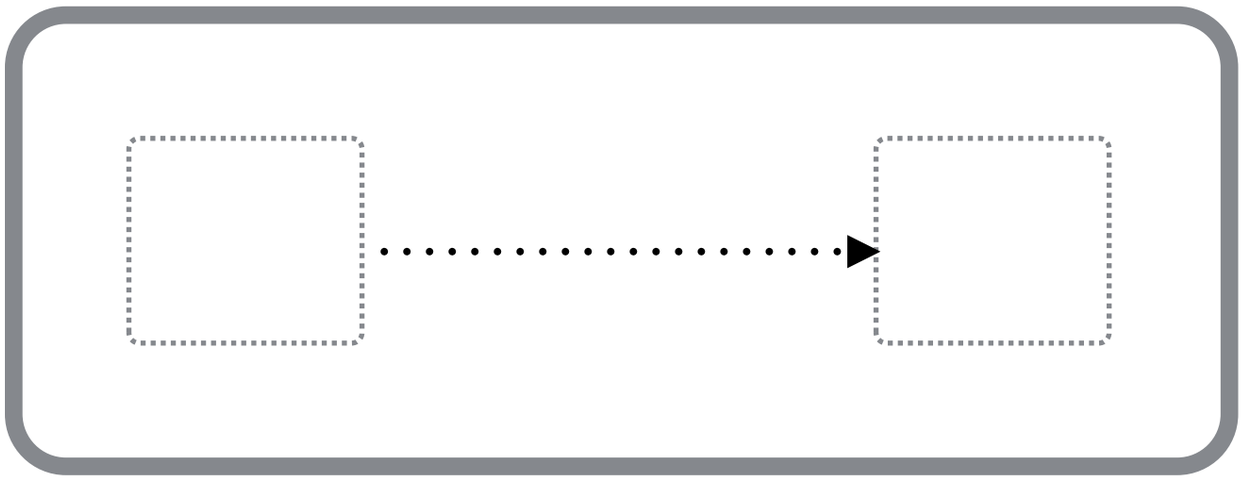


Per-call overhead is very low

Movement of data by reference

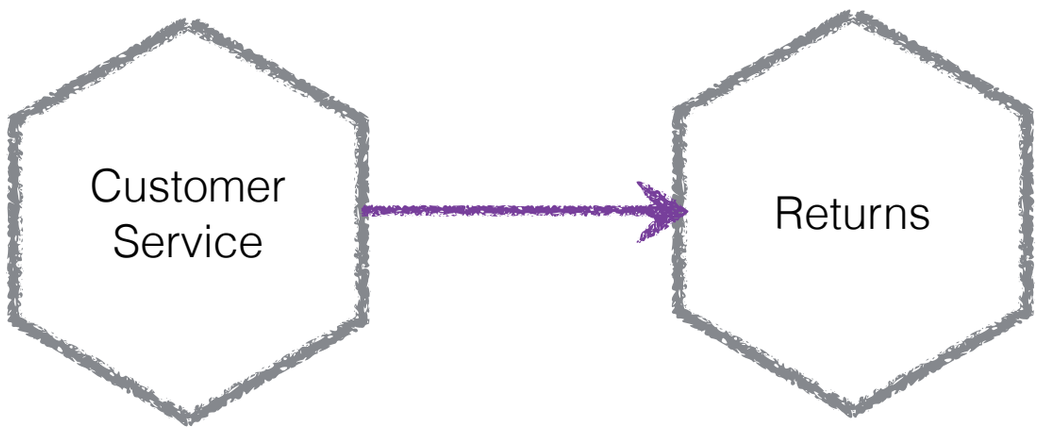


PERFORMANCE IMPLICATIONS



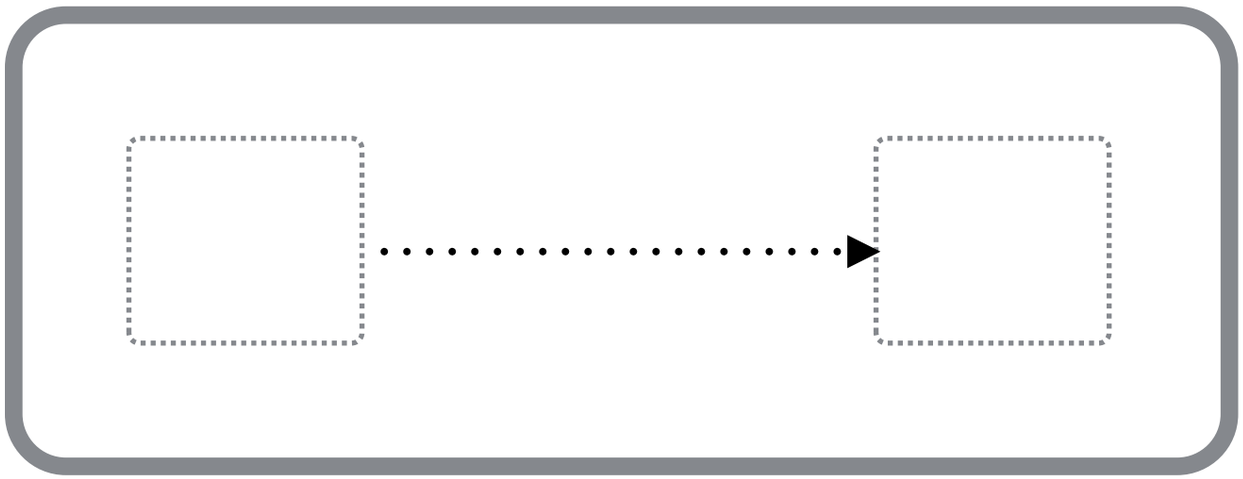
Per-call overhead is very low

Movement of data by reference



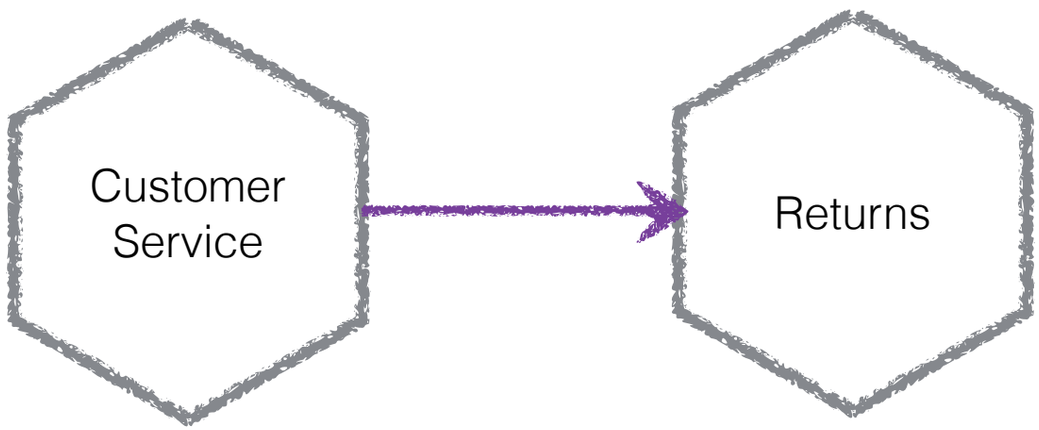
Call overhead can be very high

PERFORMANCE IMPLICATIONS



Per-call overhead is very low

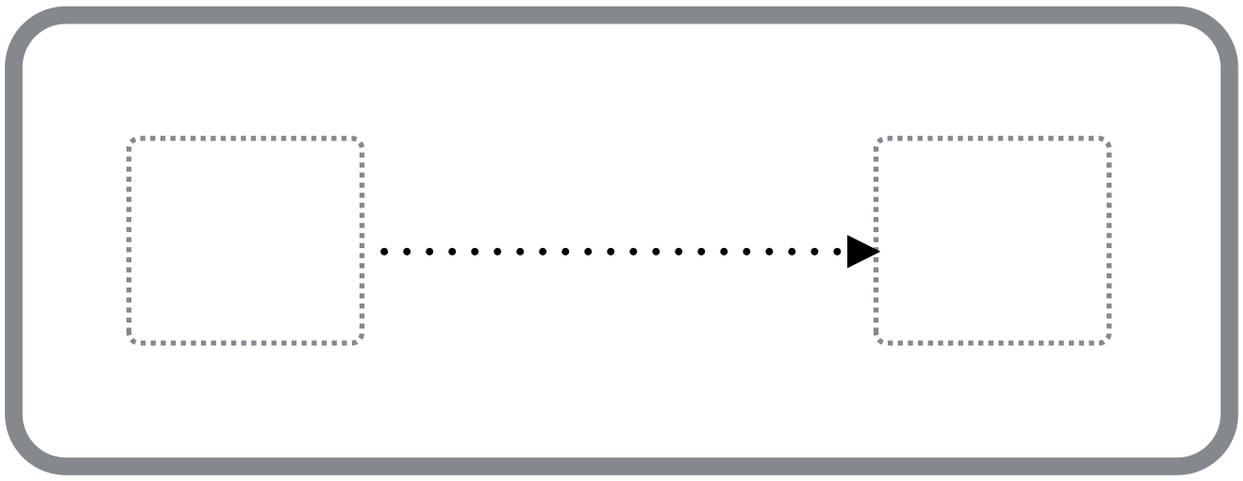
Movement of data by reference

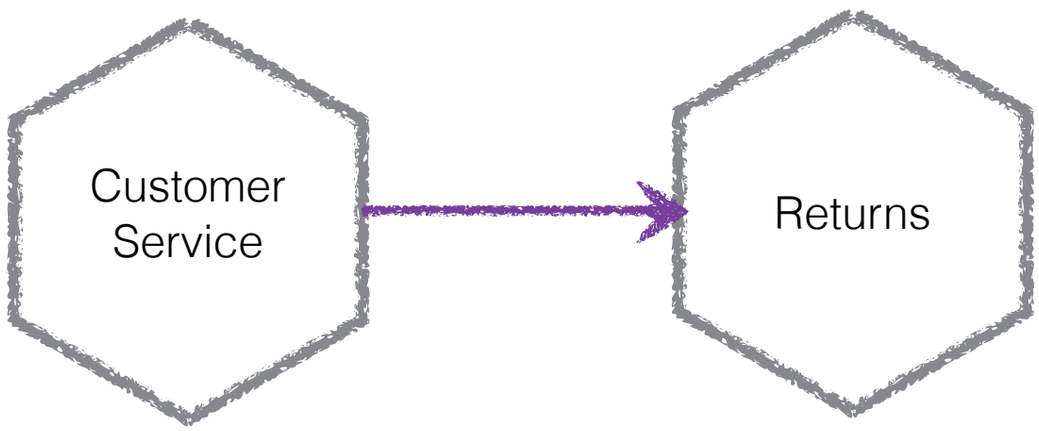


Call overhead can be very high

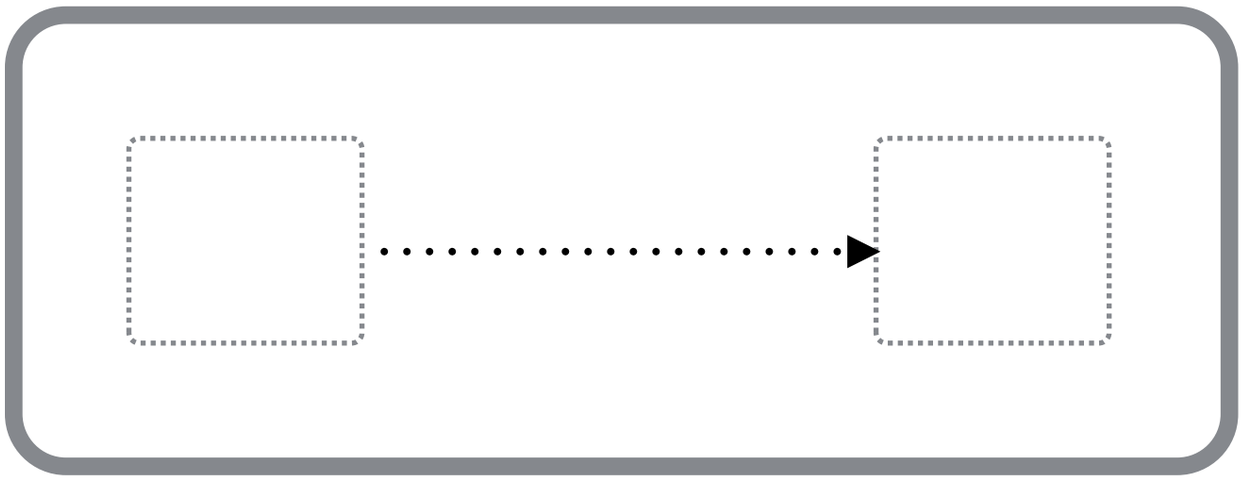
Data moved by marshalling, or handing off to an external store

HANDLING ERRORS

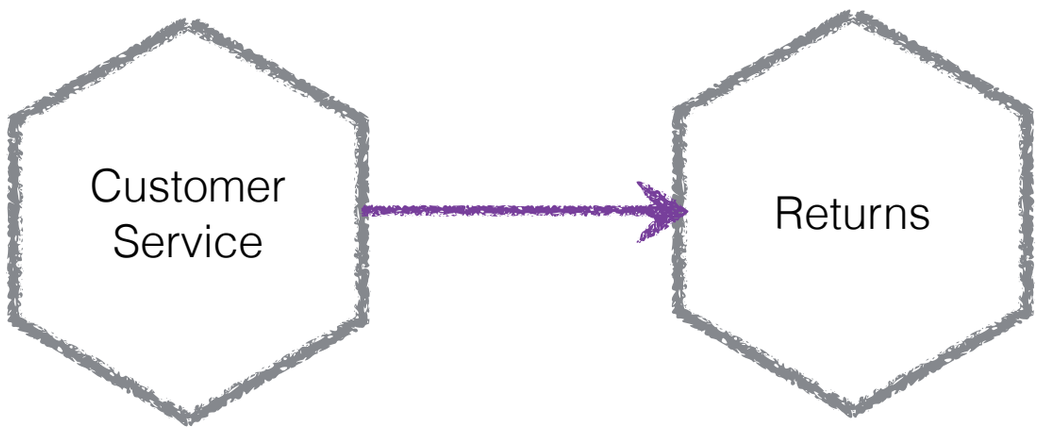




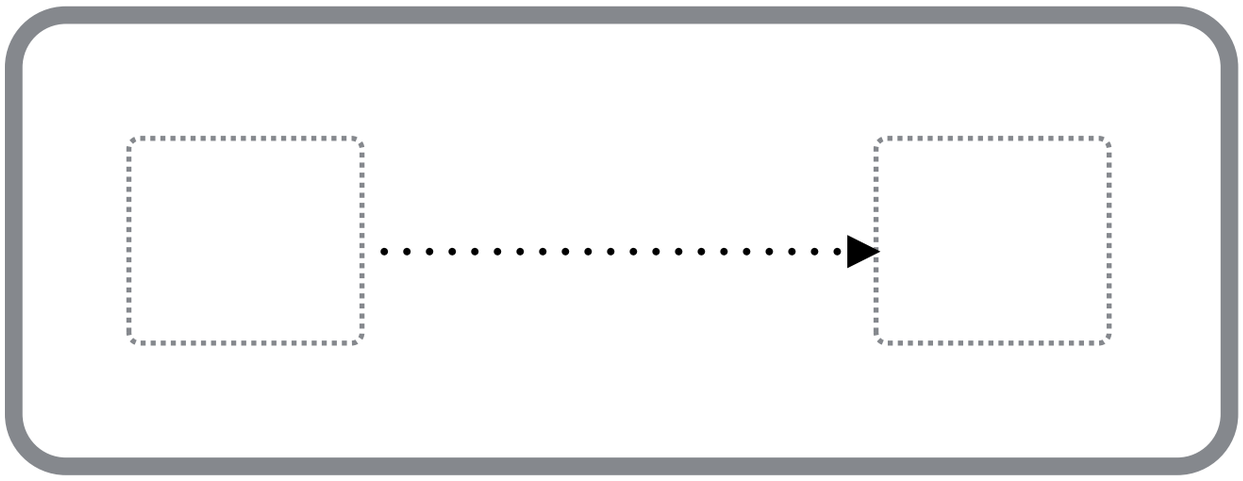
HANDLING ERRORS



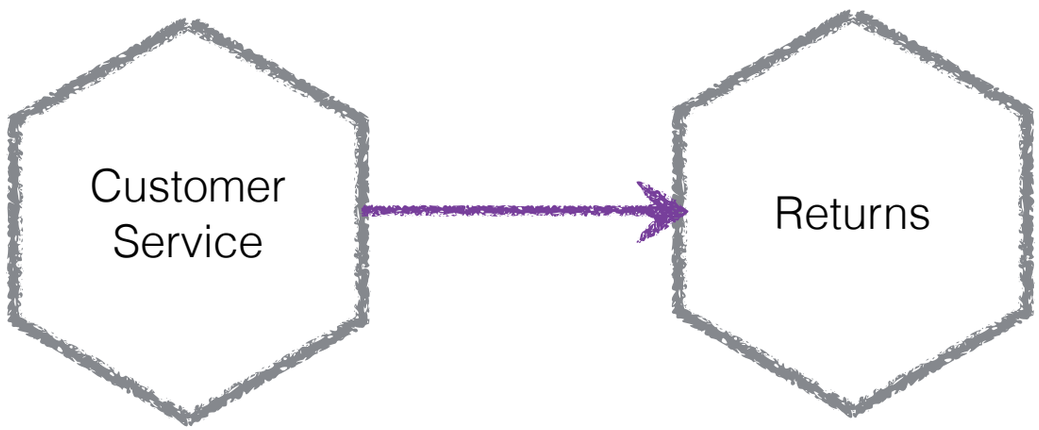
Errors straightforward



HANDLING ERRORS

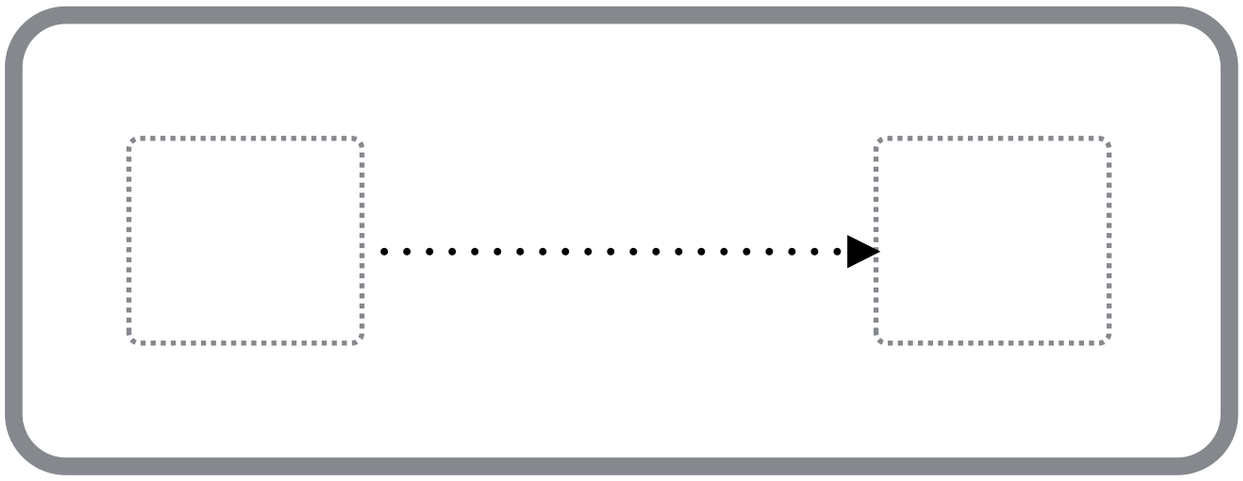


Errors straightforward

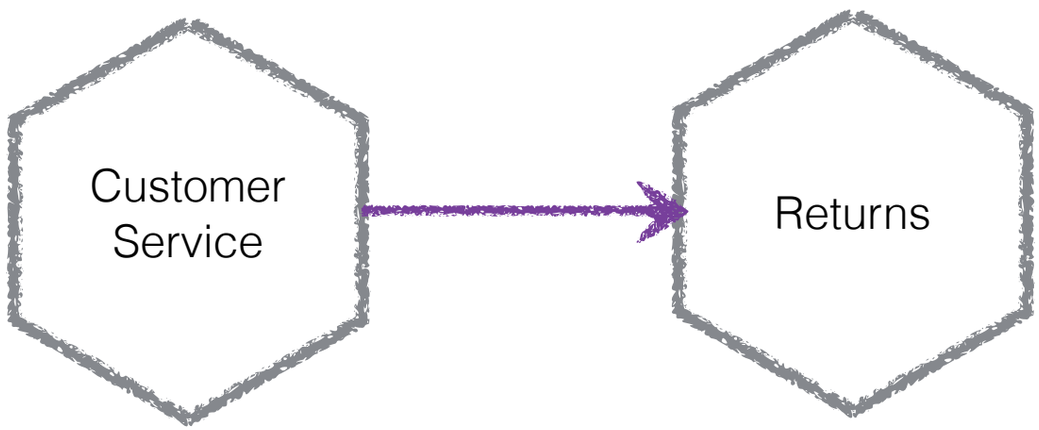


Timeouts

HANDLING ERRORS



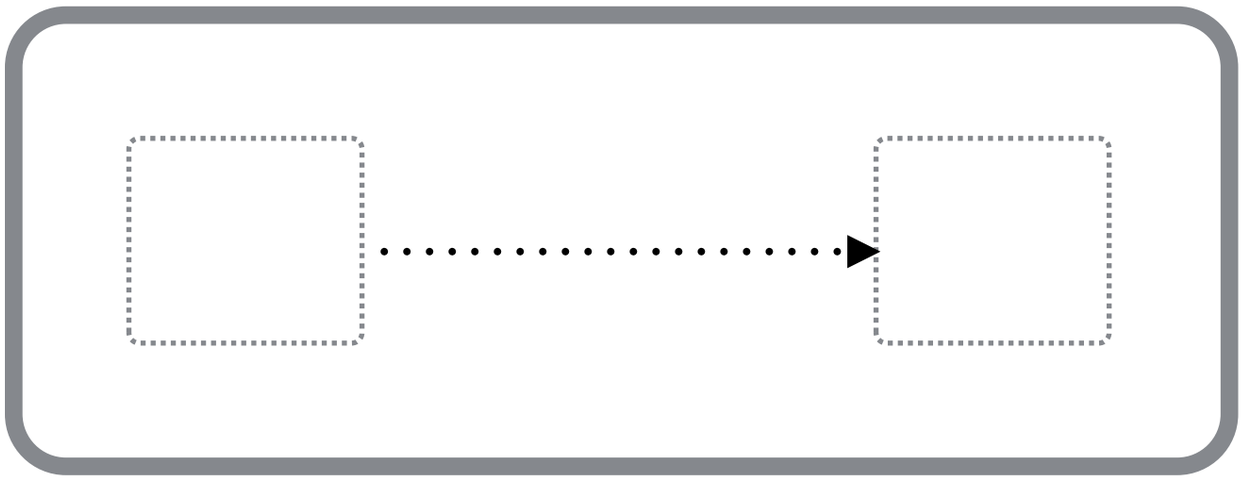
Errors straightforward



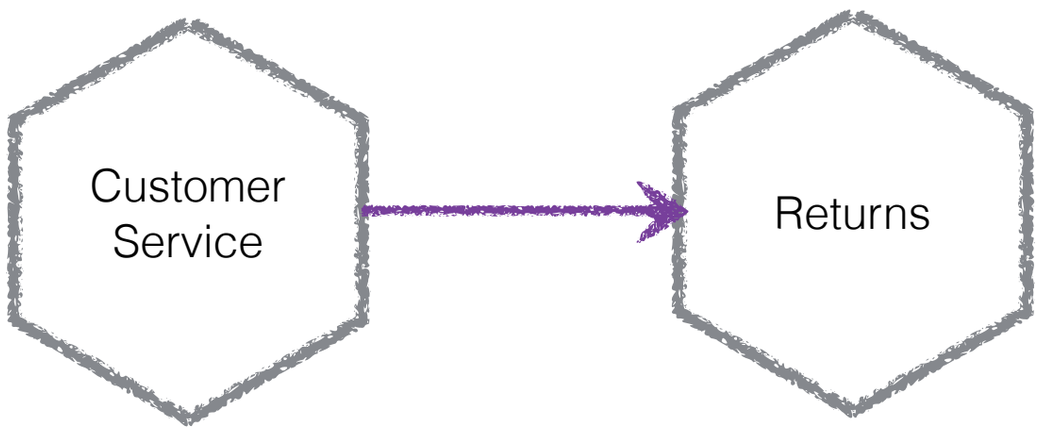
Timeouts

Downstream outage

HANDLING ERRORS



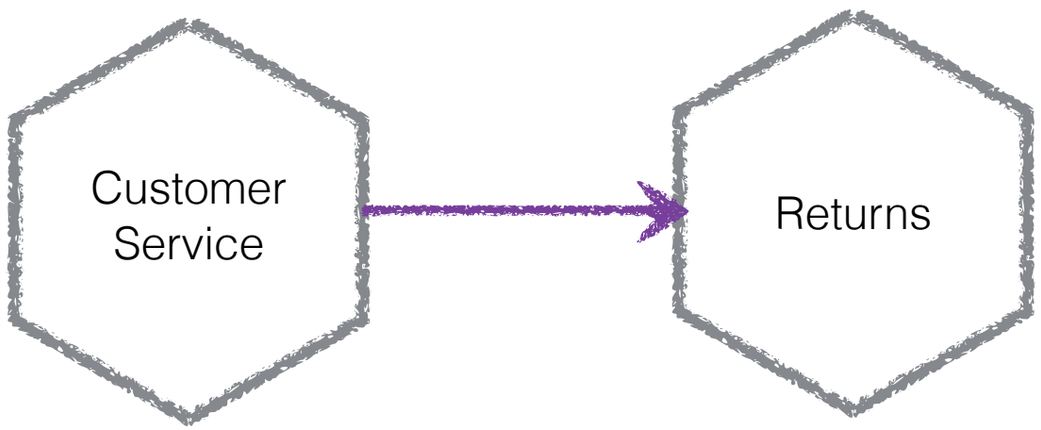
Errors straightforward



Timeouts

Downstream outage

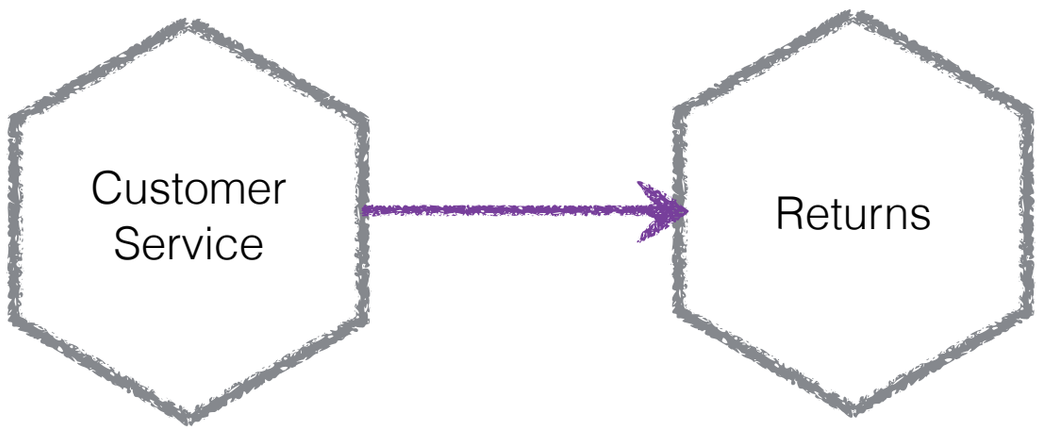
Difference between client and server errors



4XX - you did something wrong!

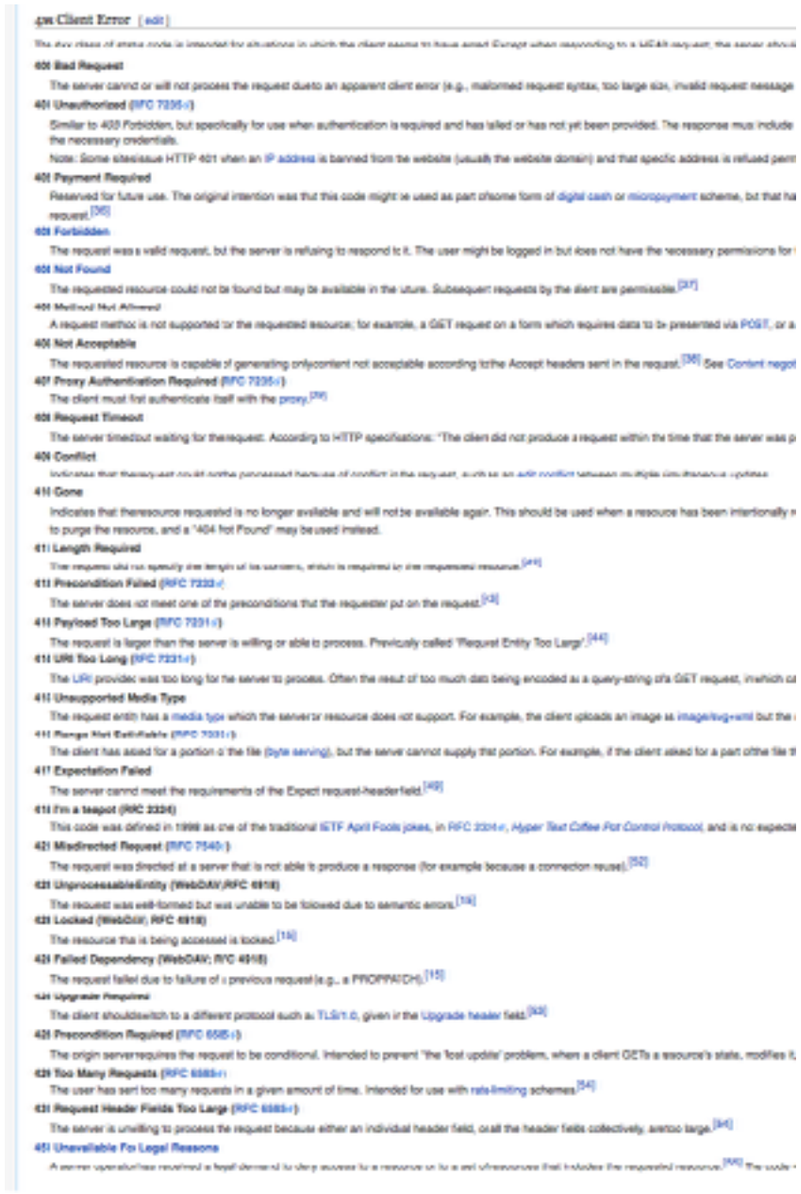
5XX - there is something wrong at my end...

4XX vs 5XX status codes



4XX - you did something wrong!

5XX - there is something wrong at my end...



4xx Client Error [edit]

The 4xx class of status code is intended for situations in which the client seems to have erred. Except when responding to a HEAD request, the server should

400 Bad Request

The server cannot or will not process the request due to an apparent client error (e.g., malformed request syntax, too large size, invalid request message

401 Unauthorized (RFC 7235)

Similar to 403 Forbidden, but specifically for use when authentication is required and has failed or has not yet been provided. The response must include the necessary credentials.

Note: Some sitesissue HTTP 401 when an IP address is banned from the website (usually the website domain) and that specific address is refused permi

402 Payment Required

Reserved for future use. The original intention was that this code might be used as part ofsome form of digital cash or micropayment scheme, but that ha request.[36]

403 Forbidden

The request was a valid request, but the server is refusing to respond to it. The user might be logged in but does not have the necessary permissions for

404 Not Found

The requested resource could not be found but may be available in the future. Subsequent requests by the client are permissible.[37]

405 Method Not Allowed

A request method is not supported for the requested resource; for example, a GET request on a form which requires data to be presented via POST, or a

406 Not Acceptable

The requested resource is capable of generating onlycontent not acceptable according tothe Accept headers sent in the request.[38] See Content negot

407 Proxy Authentication Required (RFC 7235)

The client must first authenticate itself with the proxy.[39]

408 Request Timeout

The server timedout waiting for therequest. According to HTTP specifications: "The client did not produce a request within the time that the server was p

409 Conflict

Indicates that therequest could notbe processed because of conflict in the request, such as an edit conflict between multiple simultaneous updates.

410 Gone

Indicates that theresource requested is no longer available and will not be available again. This should be used when a resource has been intentionally r to purge the resource, and a "404 Not Found" may be used instead.

411 Length Required

The request did not specify the length of its content, which is required by the requested resource.[40]

412 Precondition Failed (RFC 7232)

The server does not meet one of the preconditions that the requester put on the request.[13]

413 Payload Too Large (RFC 7231)

The request is larger than the server is willing or able to process. Previously called "Request Entity Too Large".[41]

414 URI Too Long (RFC 7231)

The URI provided was too long for the server to process. Often the result of too much data being encoded as a query-string ofa GET request, inwhich ca

415 Unsupported Media Type

The request entity has a media type which the server or resource does not support. For example, the client uploads an image as image/svg+xml but the s

416 Range Not Satisfiable (RFC 7233)

The client has asked for a portion of the file (byte serving), but the server cannot supply that portion. For example, if the client asked for a part ofthe file t

417 Expectation Failed

The server cannot meet the requirements of the Expect request-header field.[42]

418 I'm a teapot (RFC 2324)

This code was defined in 1998 as one of the traditional IETF April Fools jokes, in RFC 2324, *Hyper Text Coffee Pot Control Protocol*, and is not expecte

421 Misdirected Request (RFC 7540)

The request was directed at a server that is not able to produce a response (for example because a connection reuse).[32]

422 UnprocessableEntity (WebDAV;RFC 4918)

The request was well-formed but was unable to be followed due to semantic errors.[16]

423 Locked (WebDAV; RFC 4918)

The resource that is being accessed is locked.[16]

424 Failed Dependency (WebDAV; RFC 4918)

The request failed due to failure of a previous request (e.g., a PROPPATCH).[15]

425 Upgrade Required

The client shouldswitch to a different protocol such as TLS/1.0, given in the Upgrade header field.[33]

428 Precondition Required (RFC 6585)

The origin serverrequires the request to be conditional. Intended to prevent "the 'lost update' problem, where a client GETs a resource's state, modifies it,

429 Too Many Requests (RFC 6585)

The user has sent too many requests in a given amount of time. Intended for use with rate-limiting schemes.[34]

431 Request Header Fields Too Large (RFC 6585)

The server is unwilling to process the request because either an individual header field, orall the header fields collectively, aretoo large.[34]

451 Unavailable For Legal Reasons

A server operatorhas received a legal demand to deny access to a resource or to a set ofresources that includes the requested resource.[43] The code

5xx Server Error [edit]

The server failed to fulfill an apparently valid request.[56]

Response status codes beginning with the digit "5" indicate cases in which the server is aware that it has encountered a user. These response codes are applicable to any request method.[57]

500 Internal Server Error

A generic error message, given when an unexpected condition was encountered and no more specific message is s

501 Not Implemented

The server either does not recognize the request method, or it lacks the ability to fulfill the request. Usually this impl

502 Bad Gateway

The server was acting as a gateway or proxy and received an invalid response from the upstream server.[58]

503 Service Unavailable

The server is currently unavailable (because it is overloaded or down for maintenance). Generally, this is a tempora

504 Gateway Timeout

The server was acting as a gateway or proxy and did not receive a timely response from the upstream server.[61]

505 HTTP Version Not Supported

The server does not support the HTTP protocol version used in the request.[62]

506 Variant Also Negotiates (RFC 2295)

Transparent content negotiation for the request results in a circular reference.[63]

507 Insufficient Storage (WebDAV; RFC 4918)

The server is unable to store the representation needed to complete the request.[15]

508 Loop Detected (WebDAV; RFC 5842)

The server detected an infinite loop while processing the request (sent in lieu of 208 Already Reported).

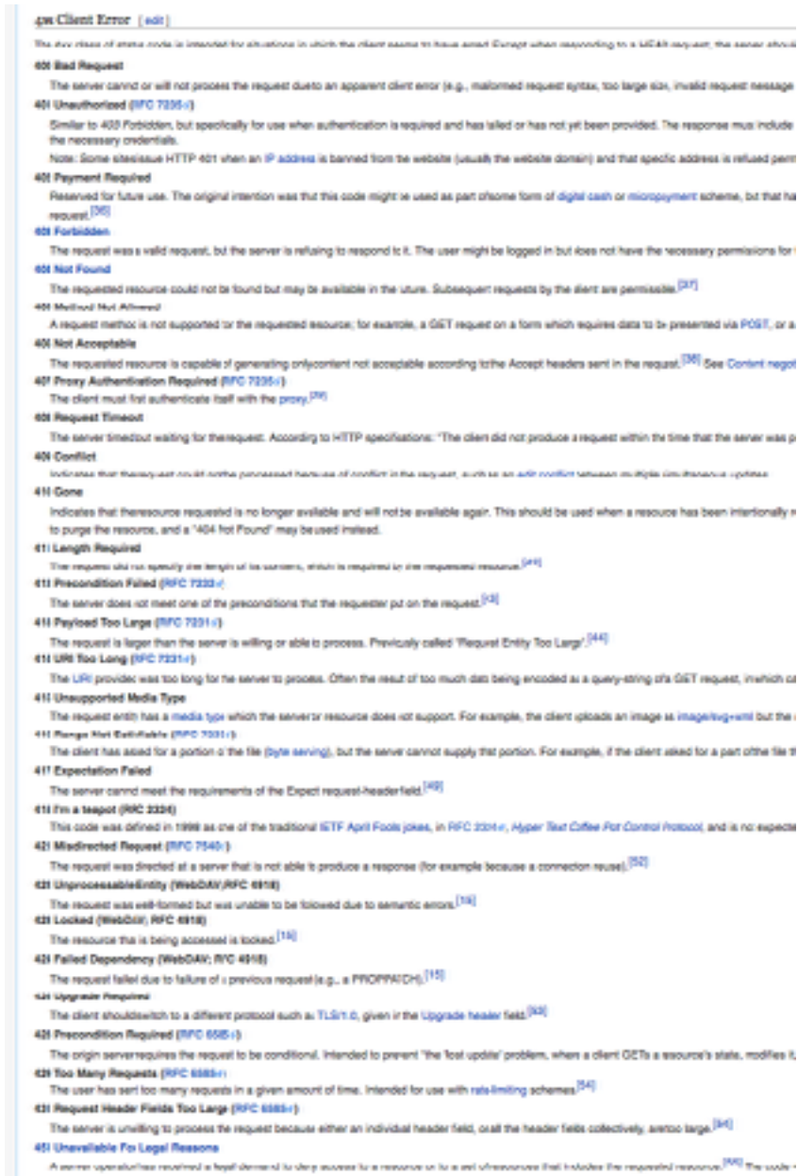
510 Not Extended (RFC 2774)

Further extensions to the request are required for the server to fulfill it.[64]

511 Network Authentication Required (RFC 6585)

The client needs to authenticate to gain network access. Intended for use by intercepting proxies used to contro ac

https://en.wikipedia.org/wiki/List_of_HTTP_status_codes



4xx Client Error [edit]

The 4xx class of status code is intended for situations in which the client seems to have erred. Except when responding to a HEAD request, the server should

400 Bad Request

The server cannot or will not process the request due to an apparent client error (e.g., malformed request syntax, too large size, invalid request message

401 Unauthorized (RFC 7235)

Similar to 403 Forbidden, but specifically for use when authentication is required and has failed or has not yet been provided. The response must include the necessary credentials.

Note: Some sitesissue HTTP 401 when an IP address is banned from the website (usually the website domain) and that specific address is refused permi

402 Payment Required

Reserved for future use. The original intention was that this code might be used as part ofsome form of digital cash or micropayment scheme, but that ha request.[36]

403 Forbidden

The request was a valid request, but the server is refusing to respond to it. The user might be logged in but does not have the necessary permissions for

404 Not Found

The requested resource could not be found but may be available in the future. Subsequent requests by the client are permissible.[37]

405 Method Not Allowed

A request method is not supported for the requested resource; for example, a GET request on a form which requires data to be presented via POST, or a

406 Not Acceptable

The requested resource is capable of generating onlycontent not acceptable according tothe Accept headers sent in the request.[38] See Content negot

407 Proxy Authentication Required (RFC 7235)

The client must first authenticate itself with the proxy.[39]

408 Request Timeout

The server timedout waiting for therequest. According to HTTP specifications: "The client did not produce a request within the time that the server was p

409 Conflict

Indicates that therequest could notbe processed because of conflict in the request, such as an edit conflict between multiple simultaneous updates.

410 Gone

Indicates that theresource requested is no longer available and will not be available again. This should be used when a resource has been intentionally r to purge the resource, and a "404 Not Found" may be used instead.

411 Length Required

The request did not specify the length of its content, which is required by the requested resource.[40]

412 Precondition Failed (RFC 7232)

The server does not meet one of the preconditions that the requester put on the request.[13]

413 Payload Too Large (RFC 7231)

The request is larger than the server is willing or able to process. Previously called "Request Entity Too Large".[41]

414 URI Too Long (RFC 7231)

The URI provided was too long for the server to process. Often the result of too much data being encoded as a query-string ofa GET request, inwhich ca

415 Unsupported Media Type

The request entity has a media type which the server or resource does not support. For example, the client uploads an image as image/svg+xml but the s

416 Range Not Satisfiable (RFC 7233)

The client has asked for a portion of the file (byte serving), but the server cannot supply that portion. For example, if the client asked for a part ofthe file t

417 Expectation Failed

The server cannot meet the requirements of the Expect request-headerfield.[42]

418 I'm a teapot (RFC 2324)

This code was defined in 1998 as one of the traditional IETF April Fools jokes, in RFC 2324, *Hyper Text Coffee Pot Control Protocol*, and is not expecte

421 Misdirected Request (RFC 7540)

The request was directed at a server that is not able to produce a response (for example because a connection reuse).[32]

422 UnprocessableEntity (WebDAV;RFC 4918)

The request was well-formed but was unable to be followed due to semantic errors.[16]

423 Locked (WebDAV; RFC 4918)

The resource that is being accessed is locked.[16]

424 Failed Dependency (WebDAV; RFC 4918)

The request failed due to failure of a previous request (e.g., a PROPPATCH).[15]

426 Upgrade Required

The client shouldswitch to a different protocol such as TLS/1.0, given in the Upgrade header field.[33]

428 Precondition Required (RFC 6585)

The origin serverrequires the request to be conditional. Intended to prevent "the 'lost update' problem, where a client GETs a resource's state, modifies it,

429 Too Many Requests (RFC 6585)

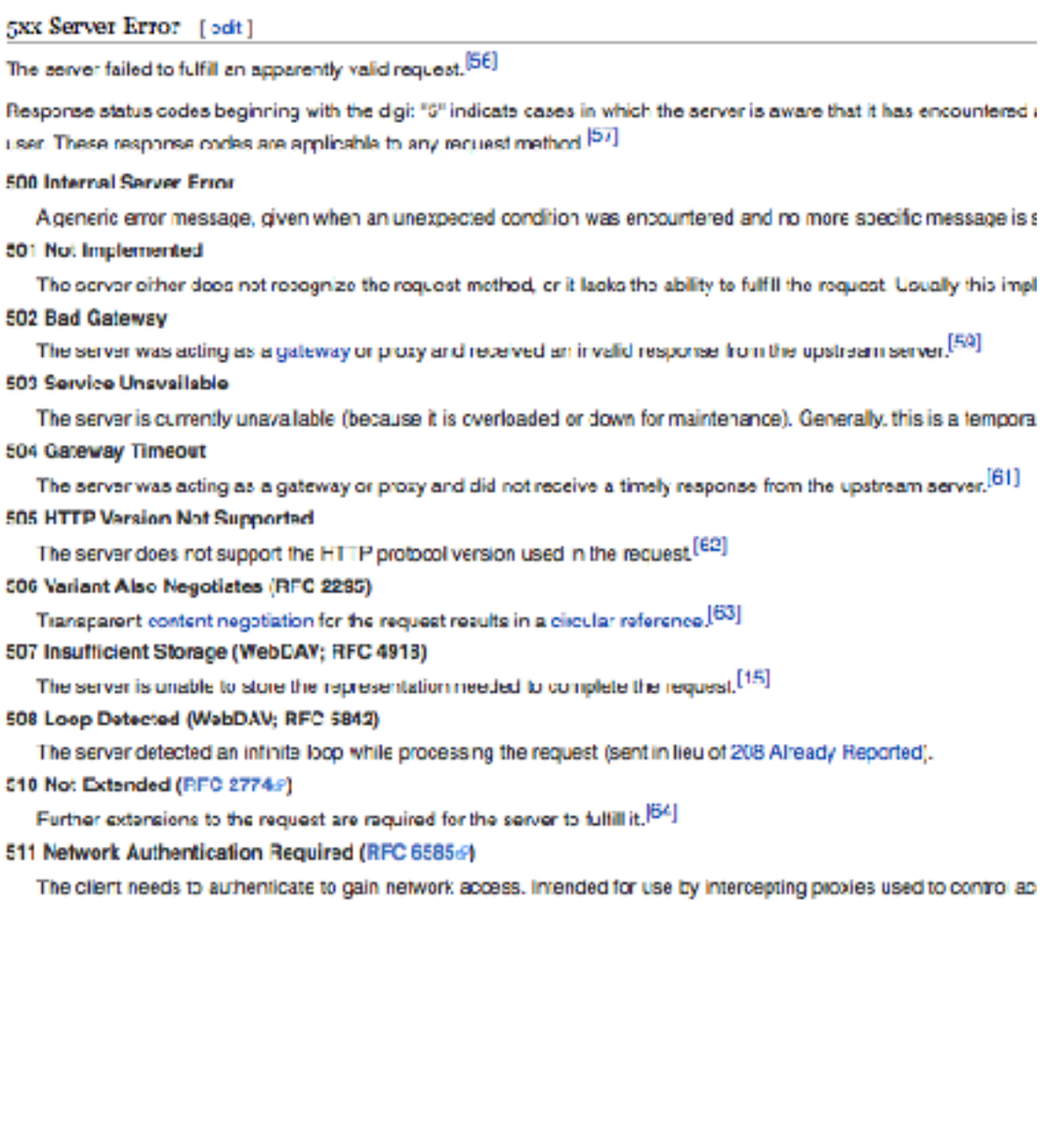
The user has sent too many requests in a given amount of time. Intended for use with rate-limiting schemes.[34]

431 Request Header Fields Too Large (RFC 6585)

The server is unwilling to process the request because either an individual header field, orall the header fields collectively, aretoo large.[34]

451 Unavailable For Legal Reasons

A server operatorhas received a legal demand to deny access to a resource or to a set ofresources that includes the requested resource.[35] The code



5xx Server Error [edit]

The server failed to fulfill an apparently valid request.[56]

Response status codes beginning with the digit "5" indicate cases in which the server is aware that it has encountered a user. These response codes are applicable to any request method.[57]

500 Internal Server Error

A generic error message, given when an unexpected condition was encountered and no more specific message is s

501 Not Implemented

The server either does not recognize the request method, or it lacks the ability to fulfill the request. Usually this impl

502 Bad Gateway

The server was acting as a gateway or proxy and received an invalid response from the upstream server.[58]

503 Service Unavailable

The server is currently unavailable (because it is overloaded or down for maintenance). Generally, this is a tempora

504 Gateway Timeout

The server was acting as a gateway or proxy and did not receive a timely response from the upstream server.[61]

505 HTTP Version Not Supported

The server does not support the HTTP protocol version used in the request.[62]

506 Variant Also Negotiates (RFC 2295)

Transparent content negotiation for the request results in a circular reference.[63]

507 Insufficient Storage (WebDAV; RFC 4918)

The server is unable to store the representation needed to complete the request.[15]

508 Loop Detected (WebDAV; RFC 5842)

The server detected an infinite loop while processing the request (sent in lieu of 208 Already Reported).

510 Not Extended (RFC 2774)

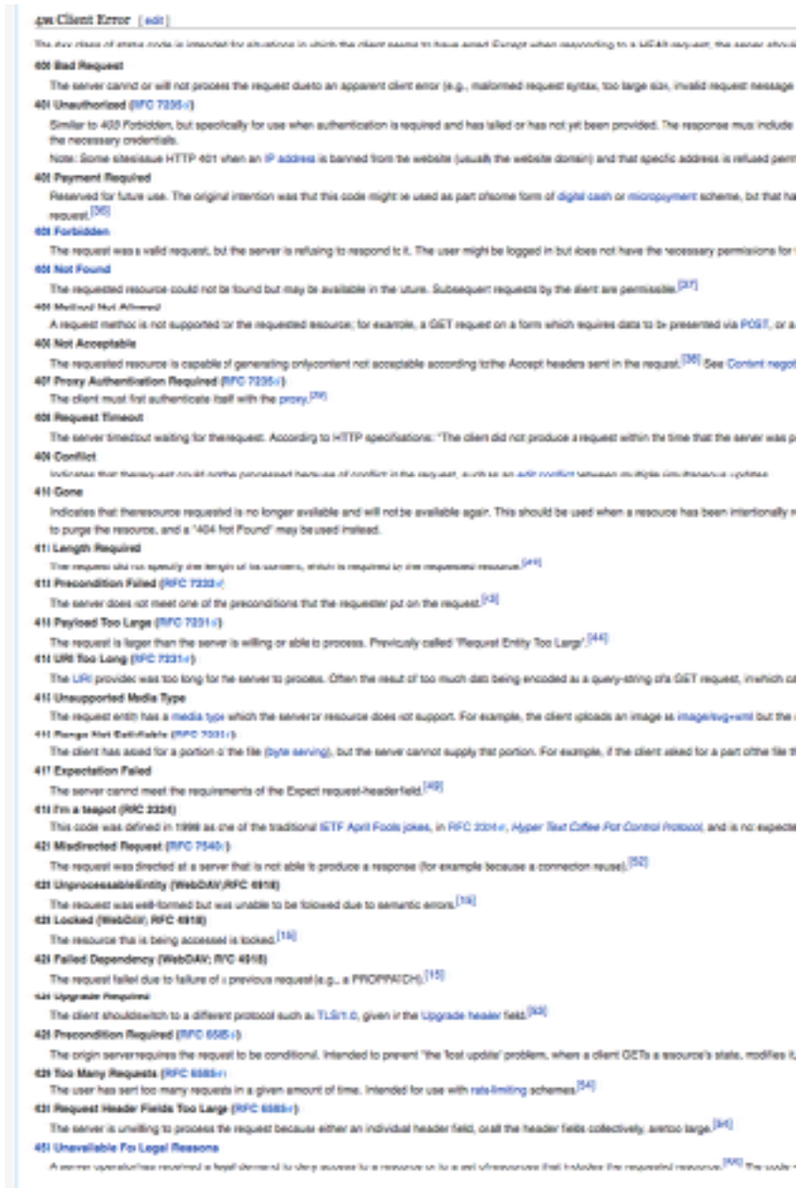
Further extensions to the request are required for the server to fulfill it.[64]

511 Network Authentication Required (RFC 6585)

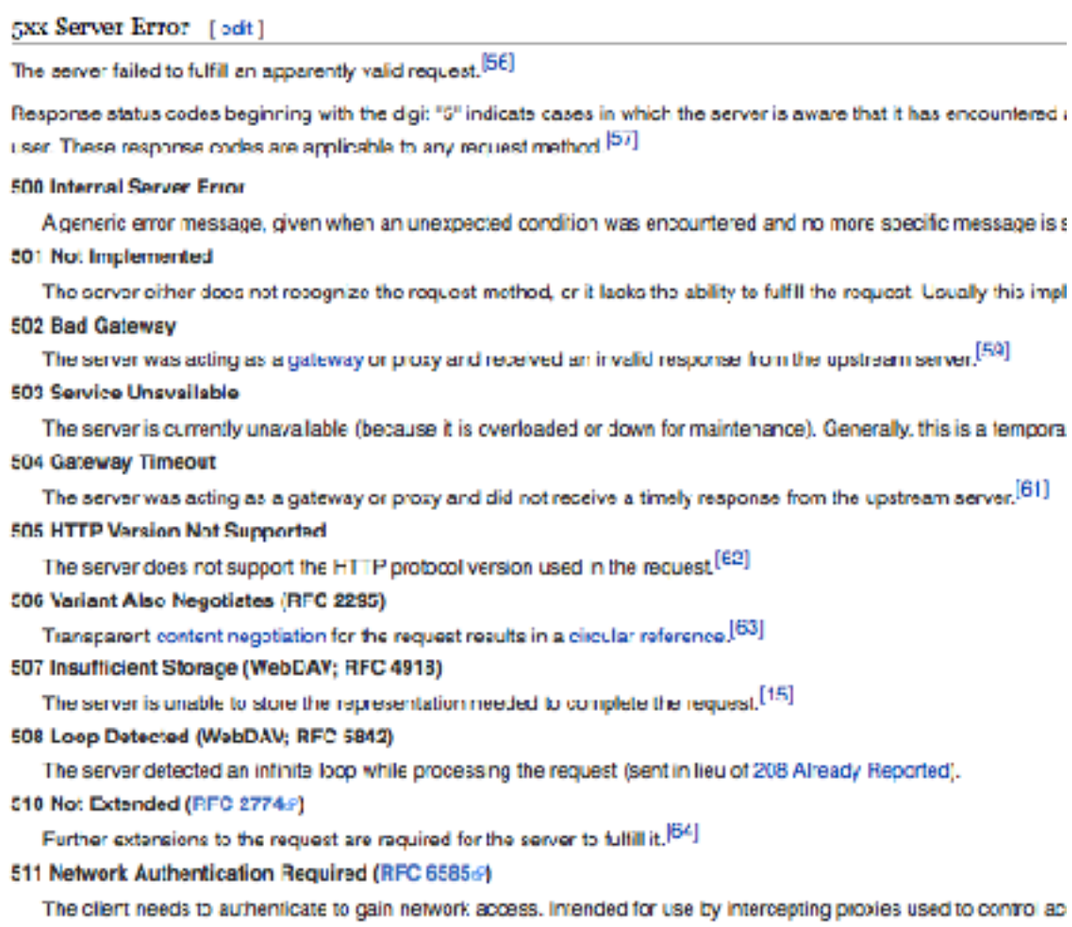
The client needs to authenticate to gain network access. Intended for use by intercepting proxies used to contro ac

28 4XX Error Codes

https://en.wikipedia.org/wiki/List_of_HTTP_status_codes



28 4XX Error Codes



11 5XX Error Codes

https://en.wikipedia.org/wiki/List_of_HTTP_status_codes

418 - I’m A Teapot

https://www.flickr.com/photos/gogllundain/1397743125/

410 - Gone

Keep it simple

Simple = Synchronous, Request Response Communication

Synchronous

Synchronous

Asynchronous

Synchronous

Block and wait

Asynchronous

Synchronous

Block and wait

Asynchronous

Fire and (maybe) forget

Synchronous

Block and wait

Asynchronous

Fire and (maybe) forget

Simple to reason about

Synchronous

Block and wait

Asynchronous

Fire and (maybe) forget

Simple to reason about

Technology more straight forward

Synchronous

Block and wait

Asynchronous

Fire and (maybe) forget

Simple to reason about

Technology more straight forward

Great for long-running jobs

Synchronous

Block and wait

Asynchronous

Fire and (maybe) forget

Simple to reason about

Technology more straight forward

Great for long-running jobs

And low latency too!

Synchronous

Block and wait

Simple to reason about

Technology more straight forward

Asynchronous

Fire and (maybe) forget

Great for long-running jobs

And low latency too!

More complex

Collaboration Styles

Collaboration Styles

Request/Response

Collaboration Styles

Request/Response

Event-based

Collaboration Styles

Request/Response

Initiate a request, expect a response

Event-based

Collaboration Styles

Request/Response

Initiate a request, expect a response

Event-based

Things happen, things react

Request/Response

Request/Response

Event-based

	Request/Response	Event-based
Synchronous		

Request/Response

Event-based

Synchronous

Asynchronous

Request/Response

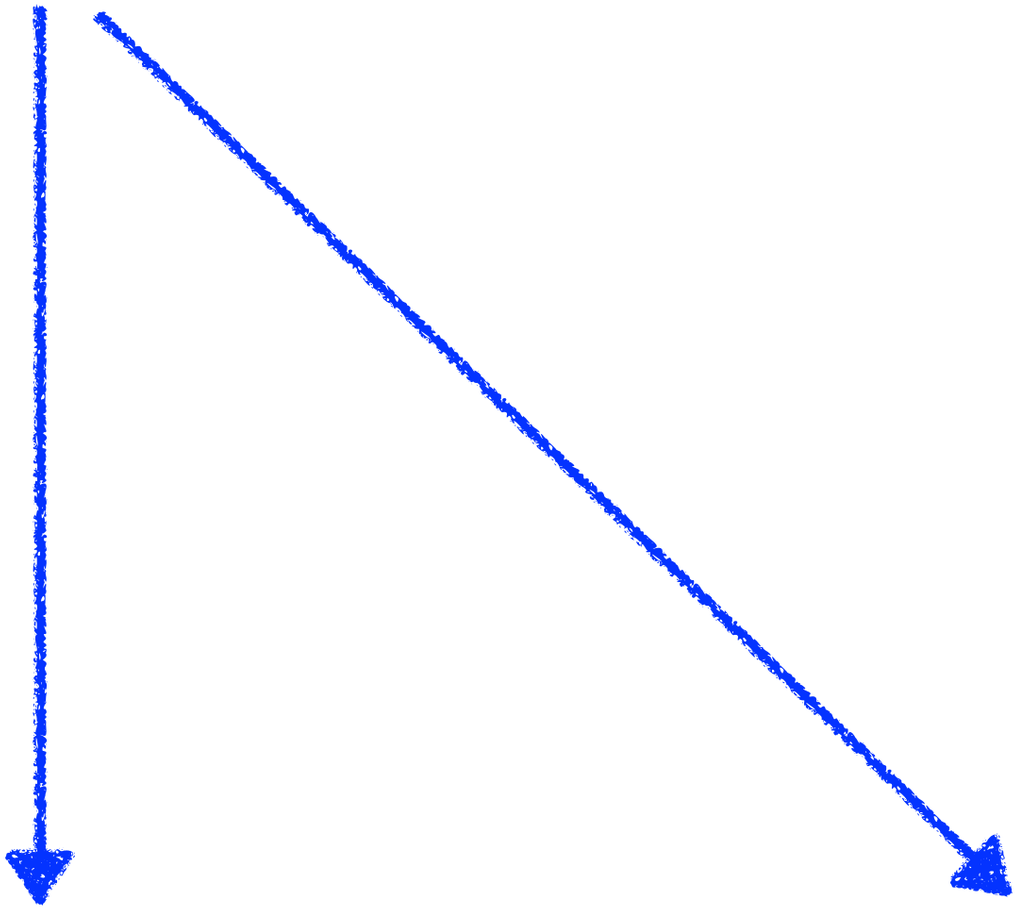
Event-based

Synchronous

Asynchronous

Request/Response

Event-based

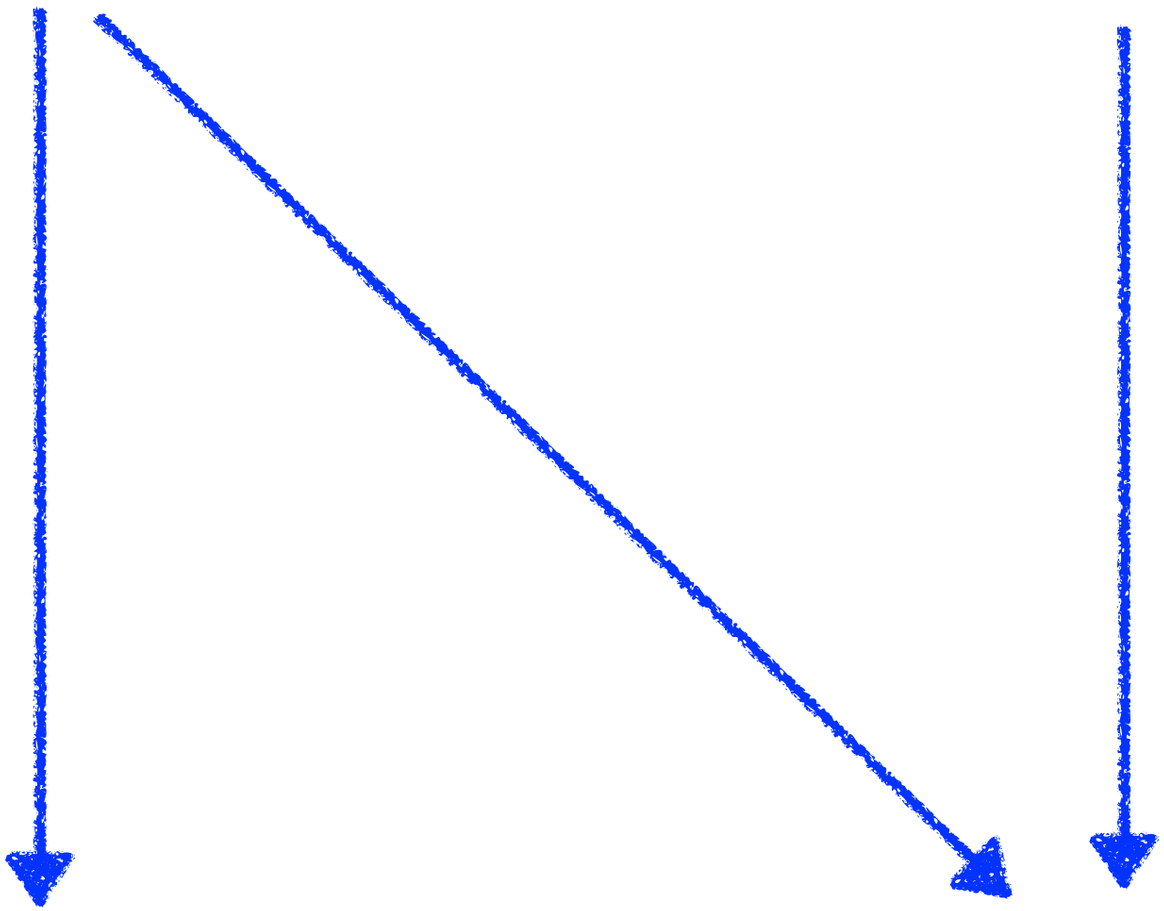


Synchronous

Asynchronous

Request/Response

Event-based



Synchronous

Asynchronous

Request/Response

Request/Response

Event-based

Request/Response

Event-based

Synchronous

	Request/Response	Event-based
Synchronous		
Asynchronous		

	Request/Response	Event-based
Synchronous	RPC HTTP	
Asynchronous		

	Request/Response	Event-based
Synchronous	RPC HTTP	
Asynchronous	RPC HTTP	

	Request/Response	Event-based
Synchronous	RPC HTTP	
Asynchronous	RPC HTTP	HTTP

	Request/Response	Event-based
Synchronous	RPC HTTP	
Asynchronous	RPC HTTP Akka	HTTP

	Request/Response	Event-based
Synchronous	RPC HTTP	
Asynchronous	RPC HTTP Akka Message Brokers	HTTP Message Brokers

Simple = Synchronous, Request Response Communication

HTTP

Very well supported

Good error handling semantics

Cache controls

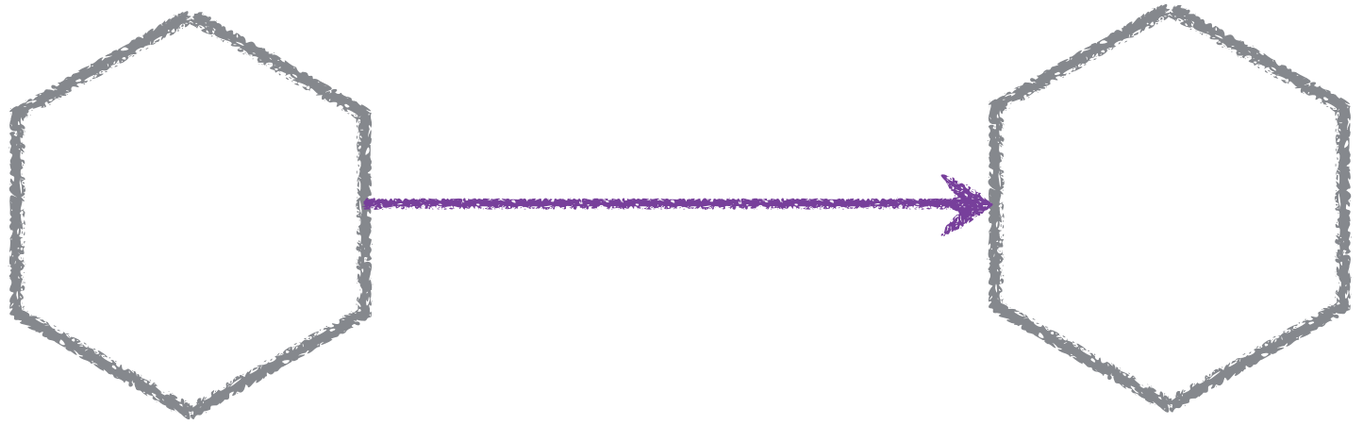
Easy to scale - good, reliable (boring?) technology

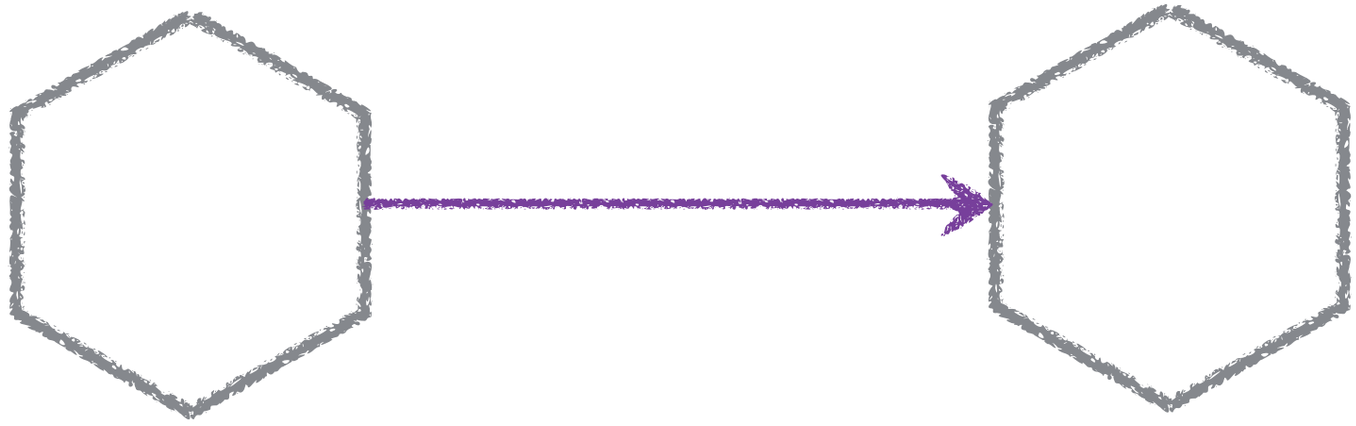
Where does this break down?

https://www.flickr.com/photos/inthe-arena/15501000164/

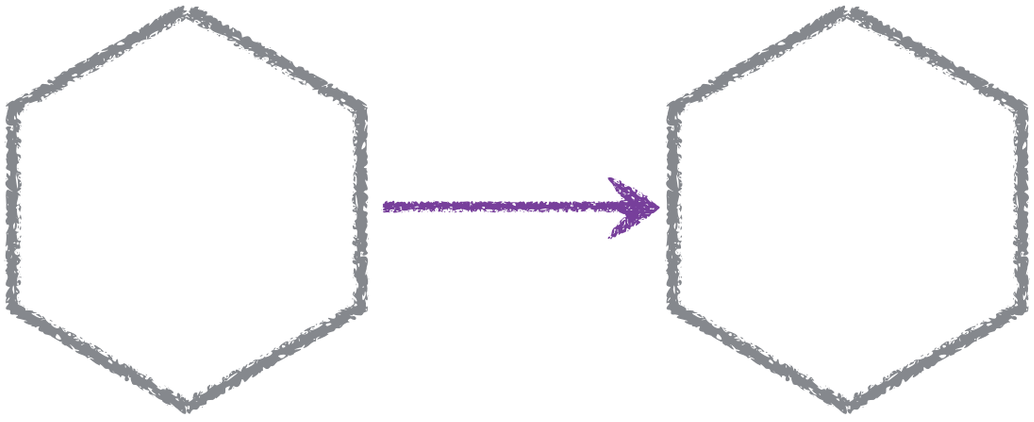
Long-lived processes

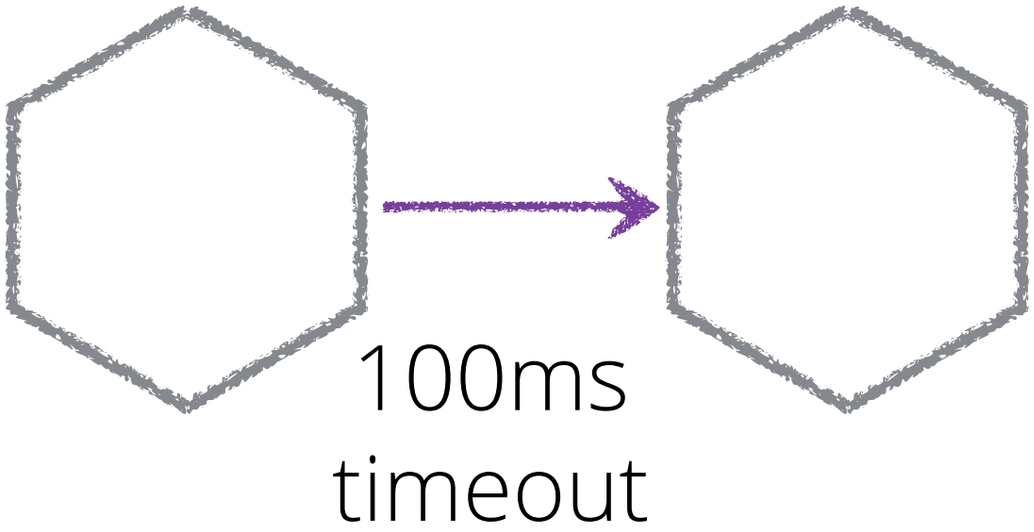
https://www.flickr.com/photos/inthe-arena/15501000164/

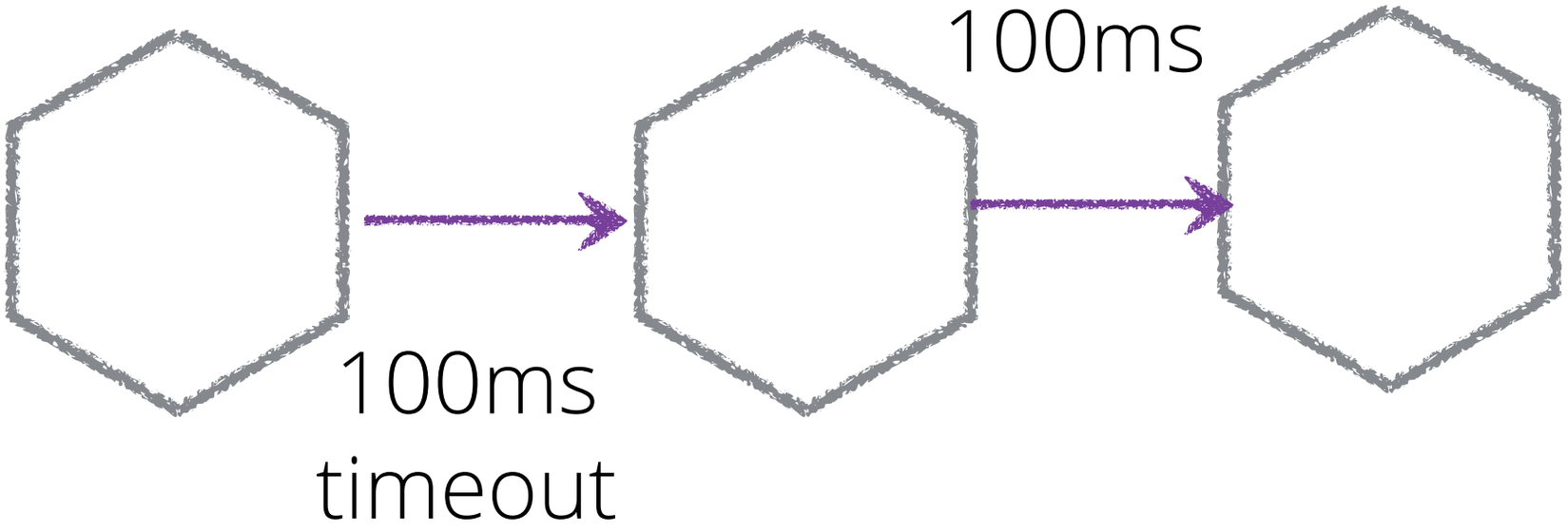


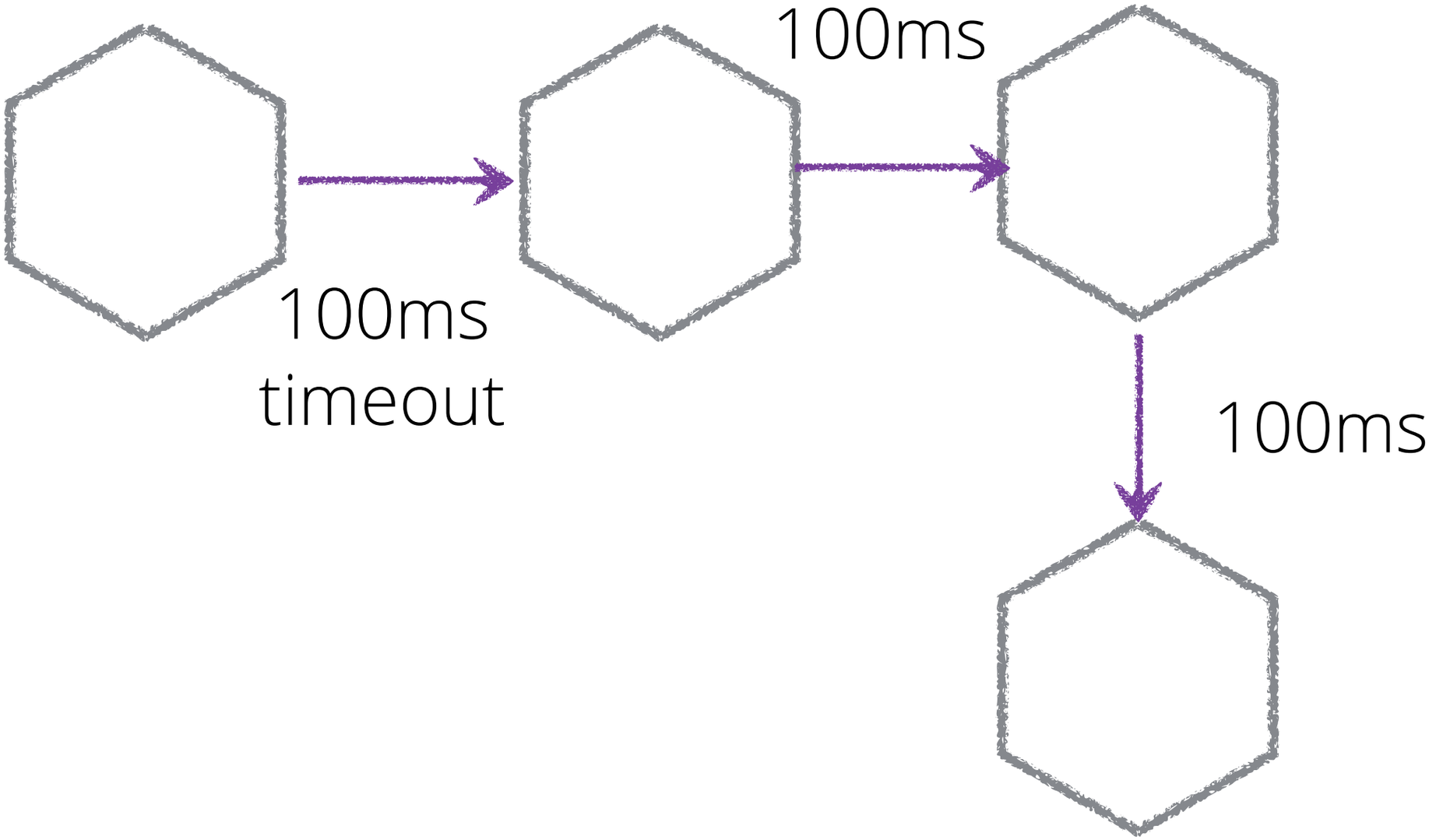


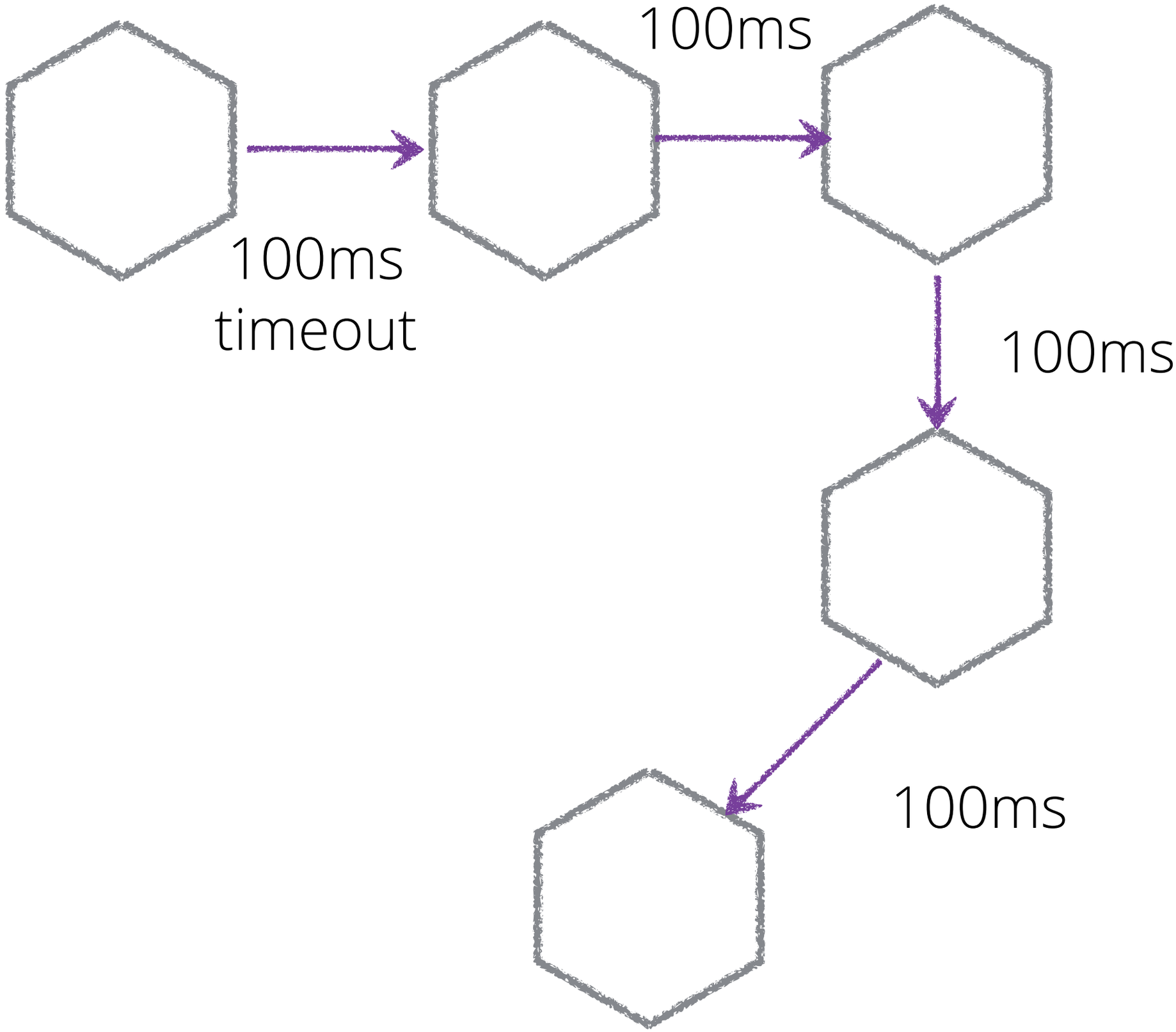
Call overhead

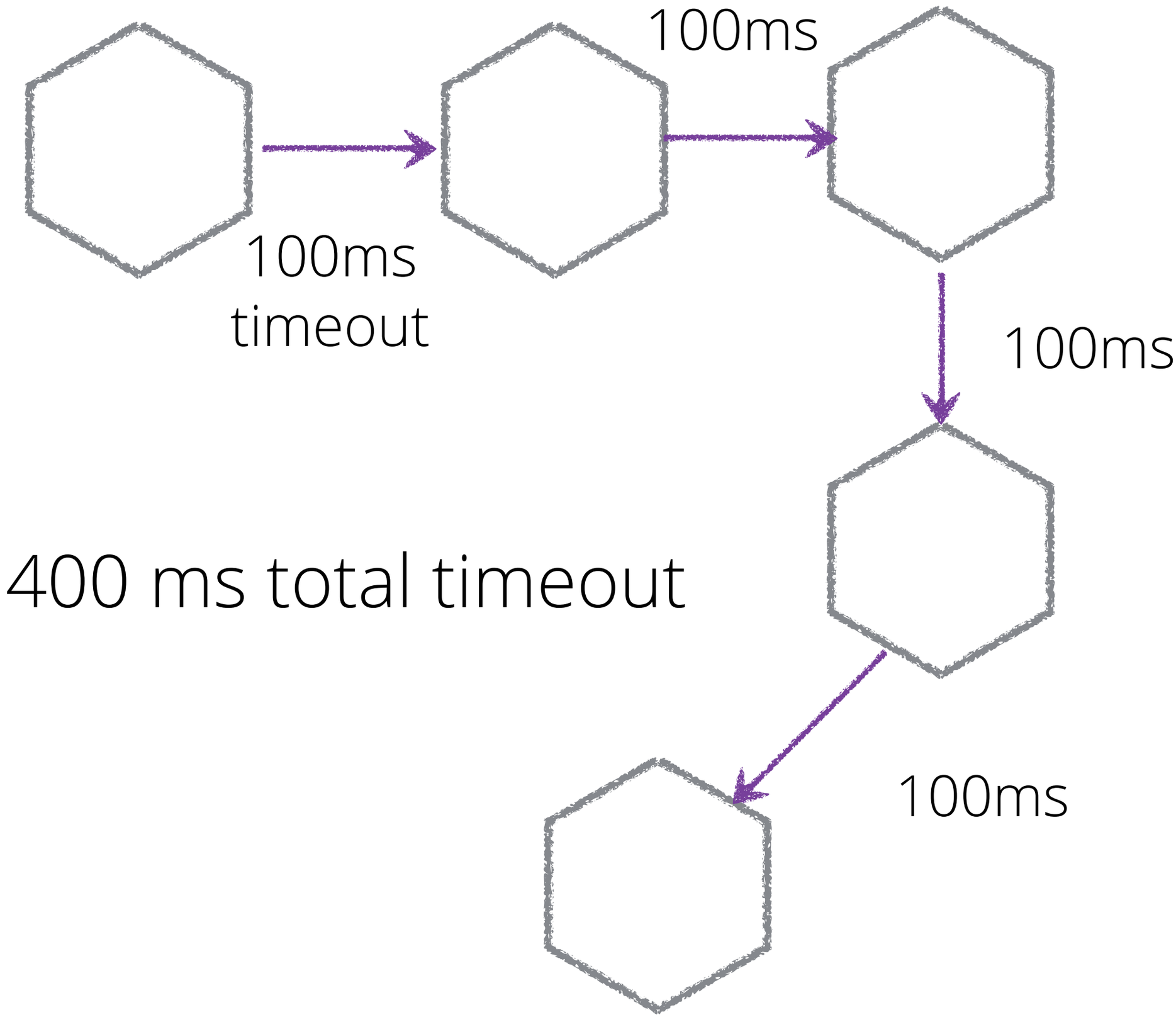



100ms
timeout


100ms
100ms
timeout

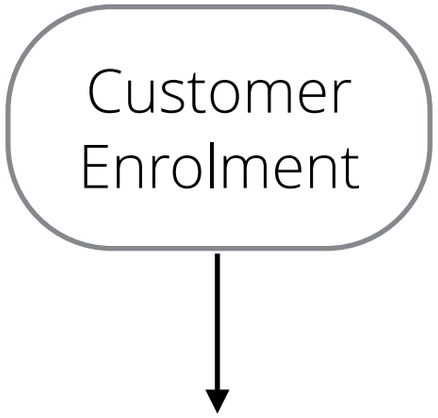

100ms
100ms
timeout
100ms

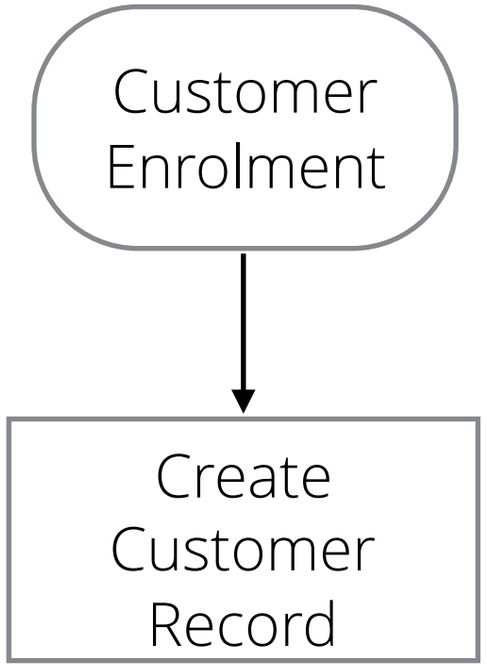

100ms
100ms
timeout
100ms
100ms

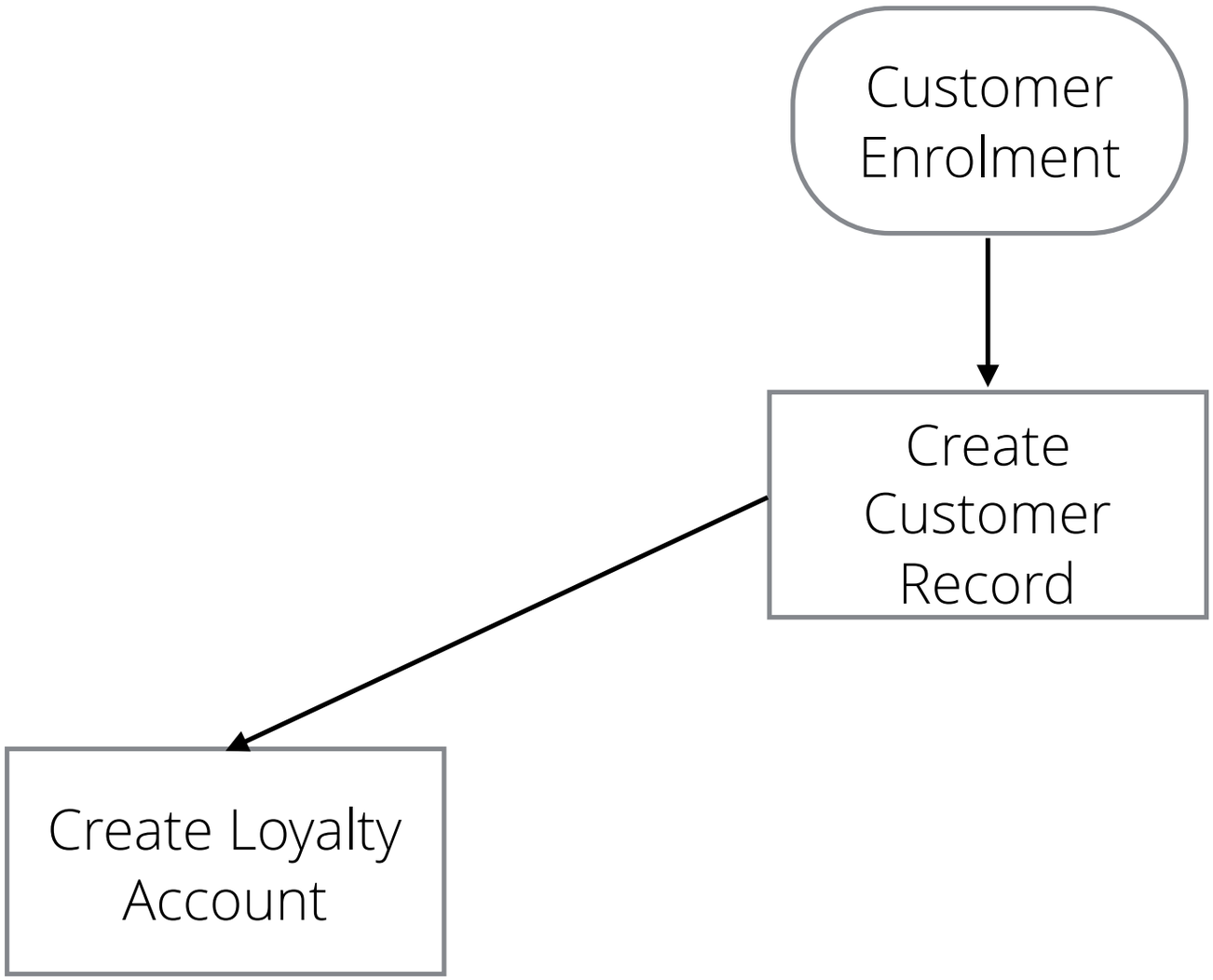


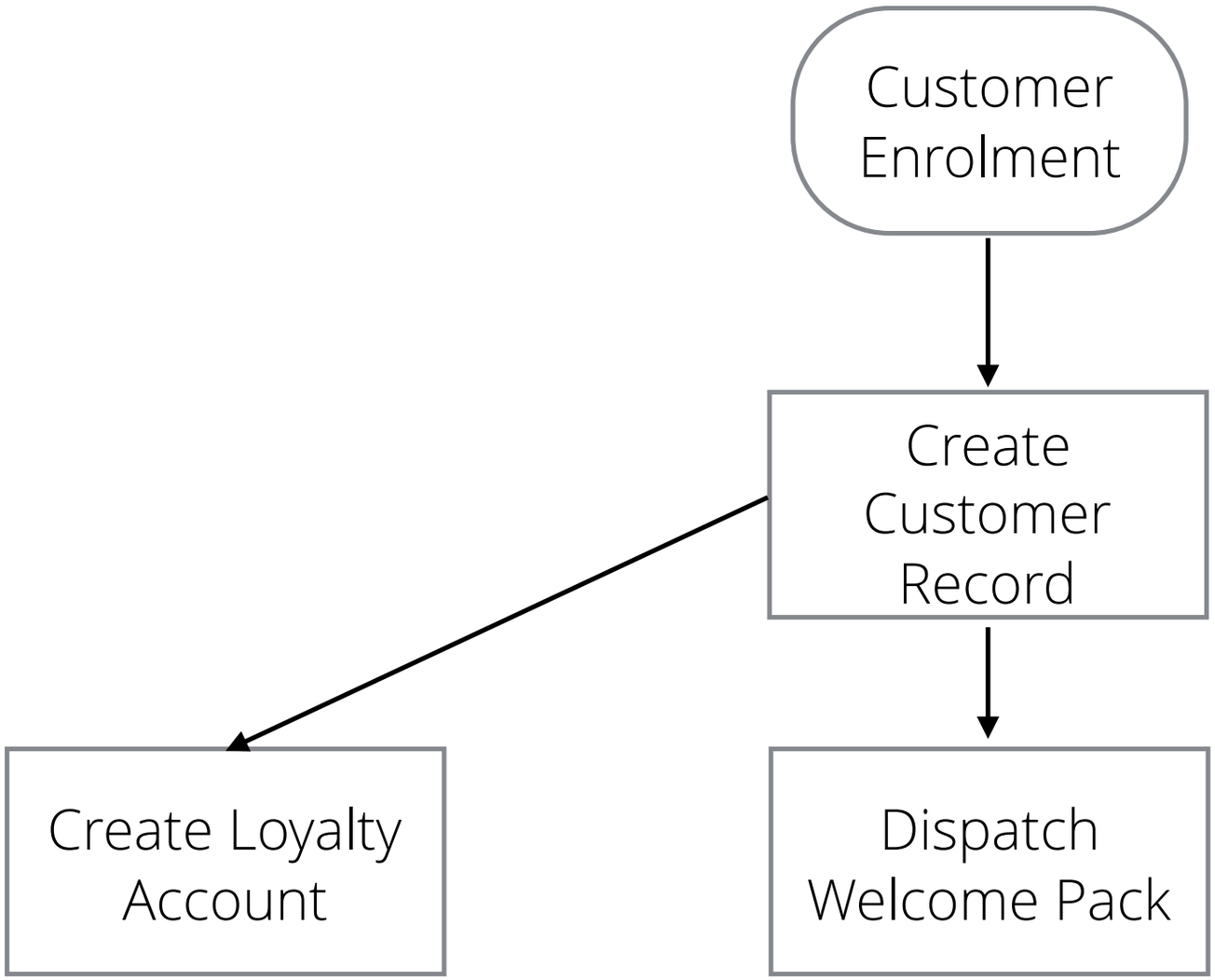
400 ms total timeout

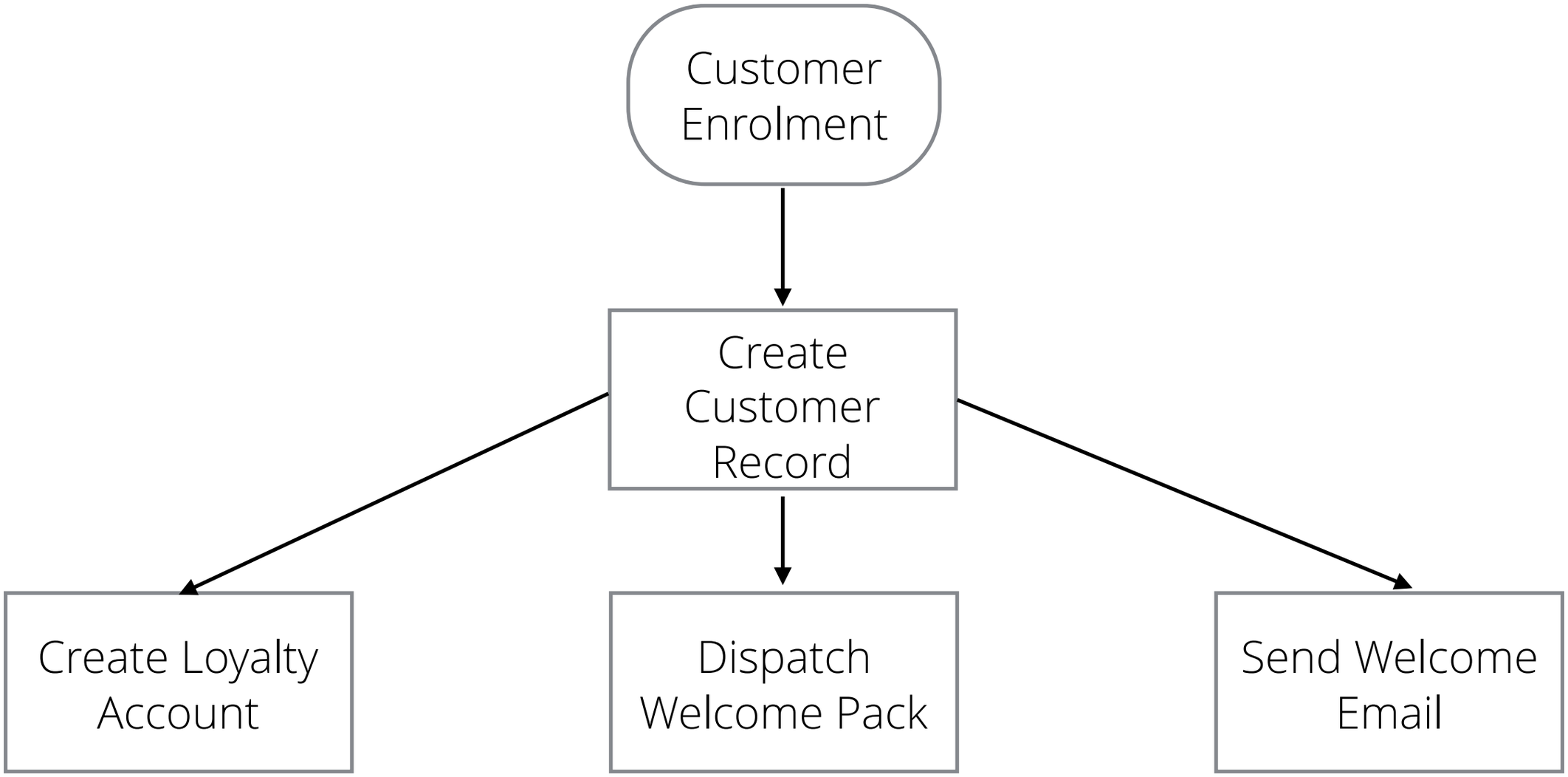
Choreography vs Orchestration

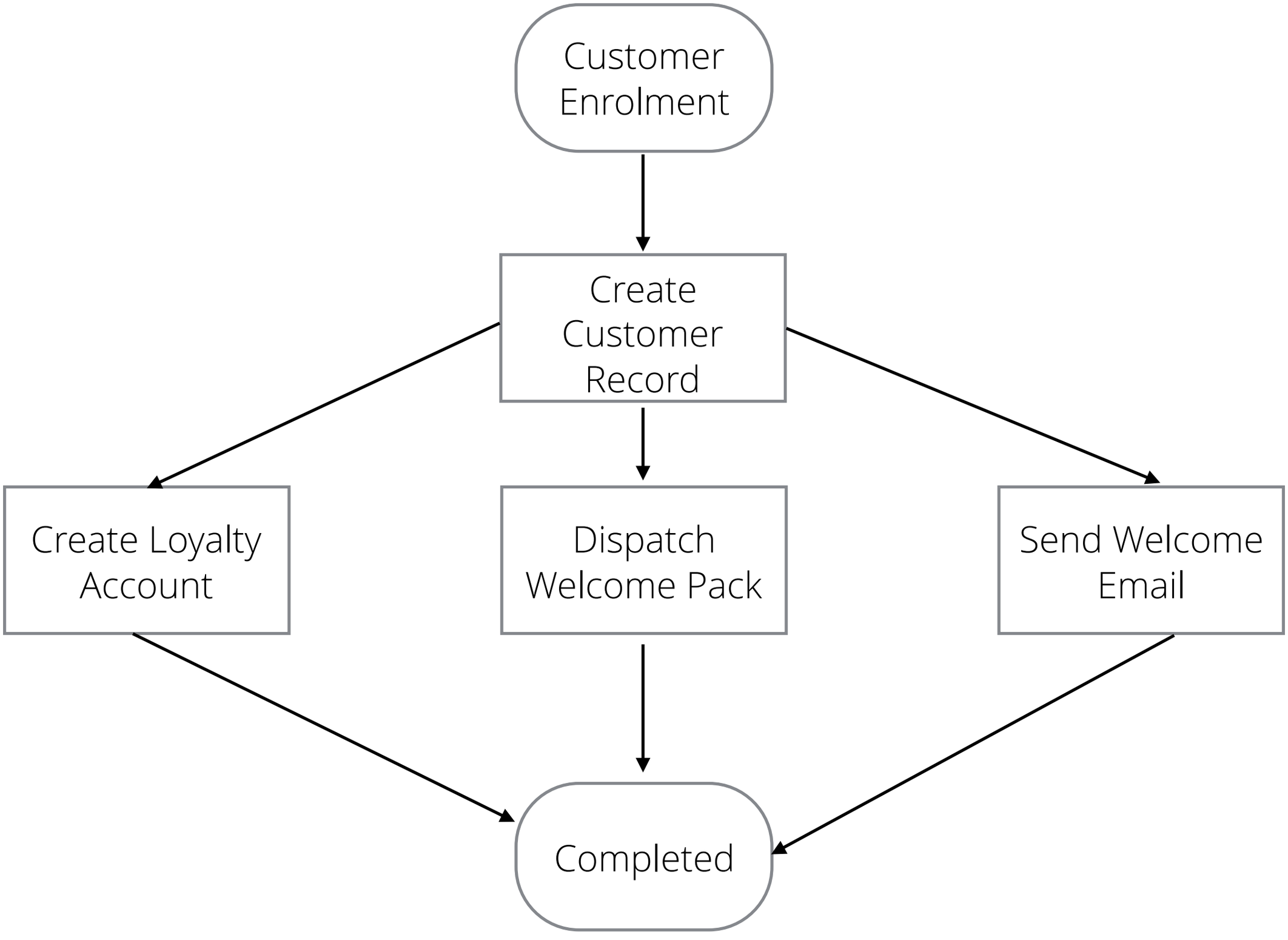

Customer
Enrolment


Customer Enrolment
Create Customer Record


Customer Enrolment
Create Customer Record
Create Loyalty Account

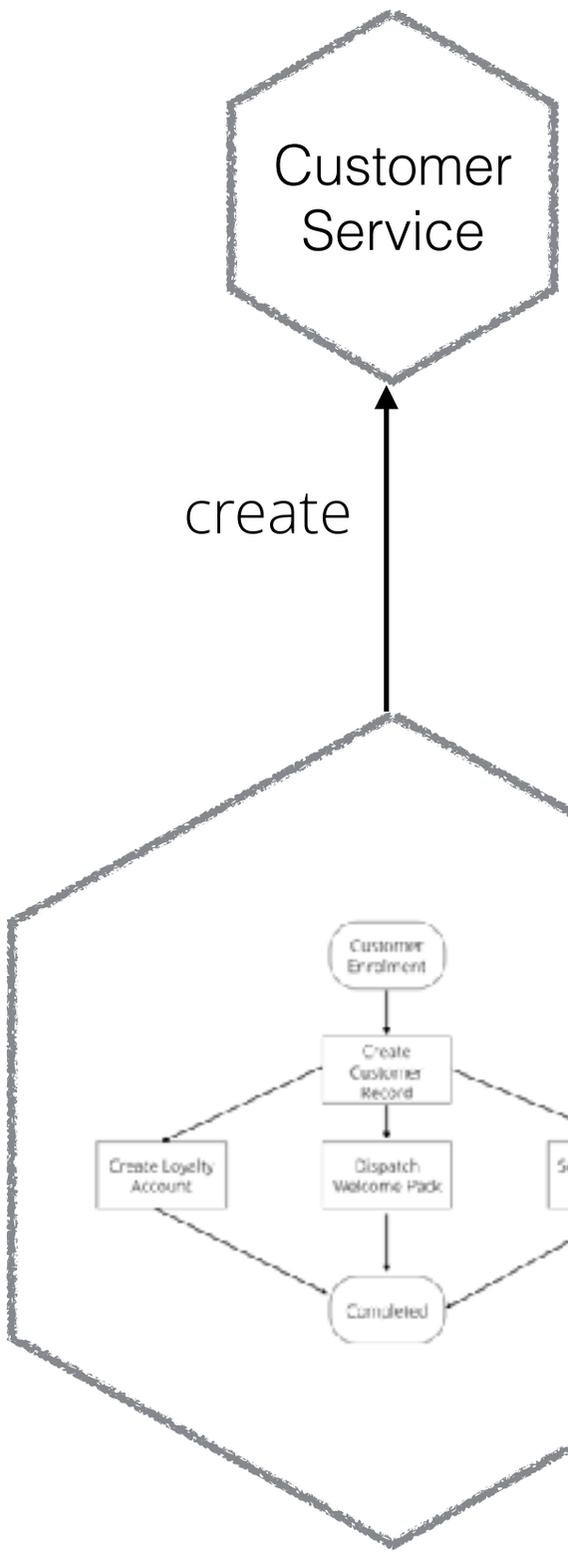

Customer Enrolment
Create Customer Record
Create Loyalty Account
Dispatch Welcome Pack


Customer Enrolment
Create Customer Record
Create Loyalty Account
Dispatch Welcome Pack
Send Welcome Email

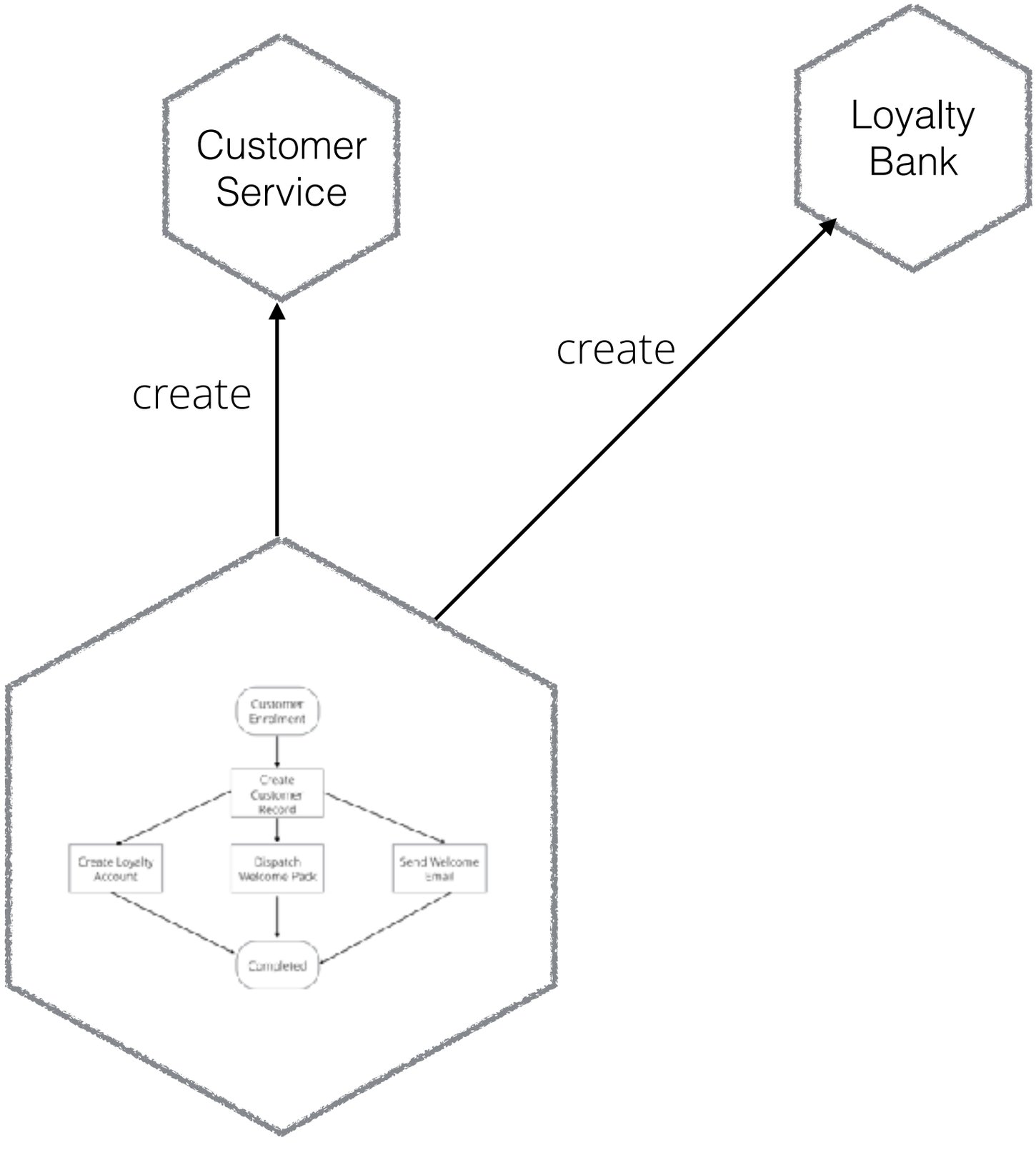

Customer Enrolment
Create Customer Record
Create Loyalty Account
Dispatch Welcome Pack
Send Welcome Email
Completed

ORCHESTRATION

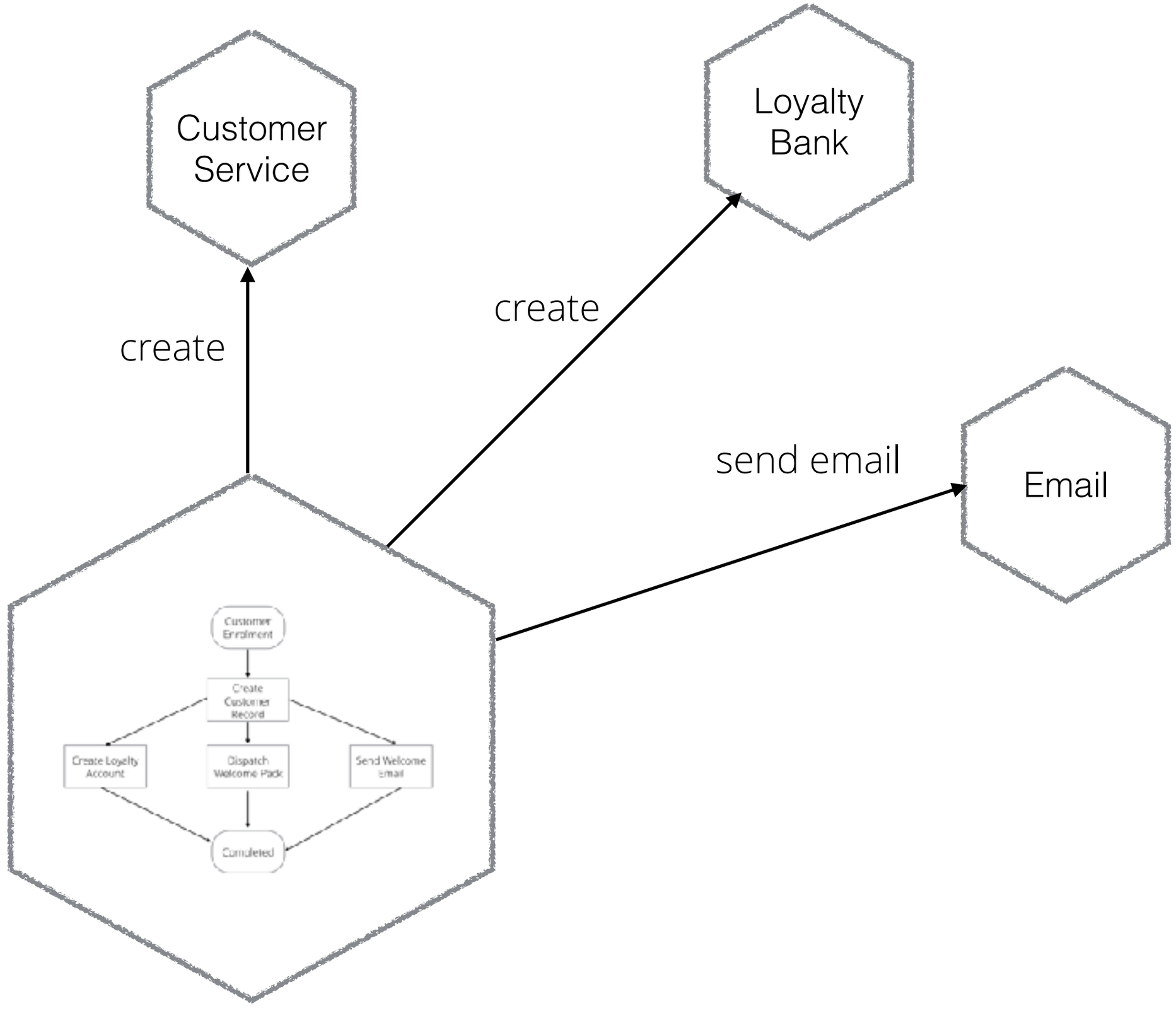
ORCHESTRATION



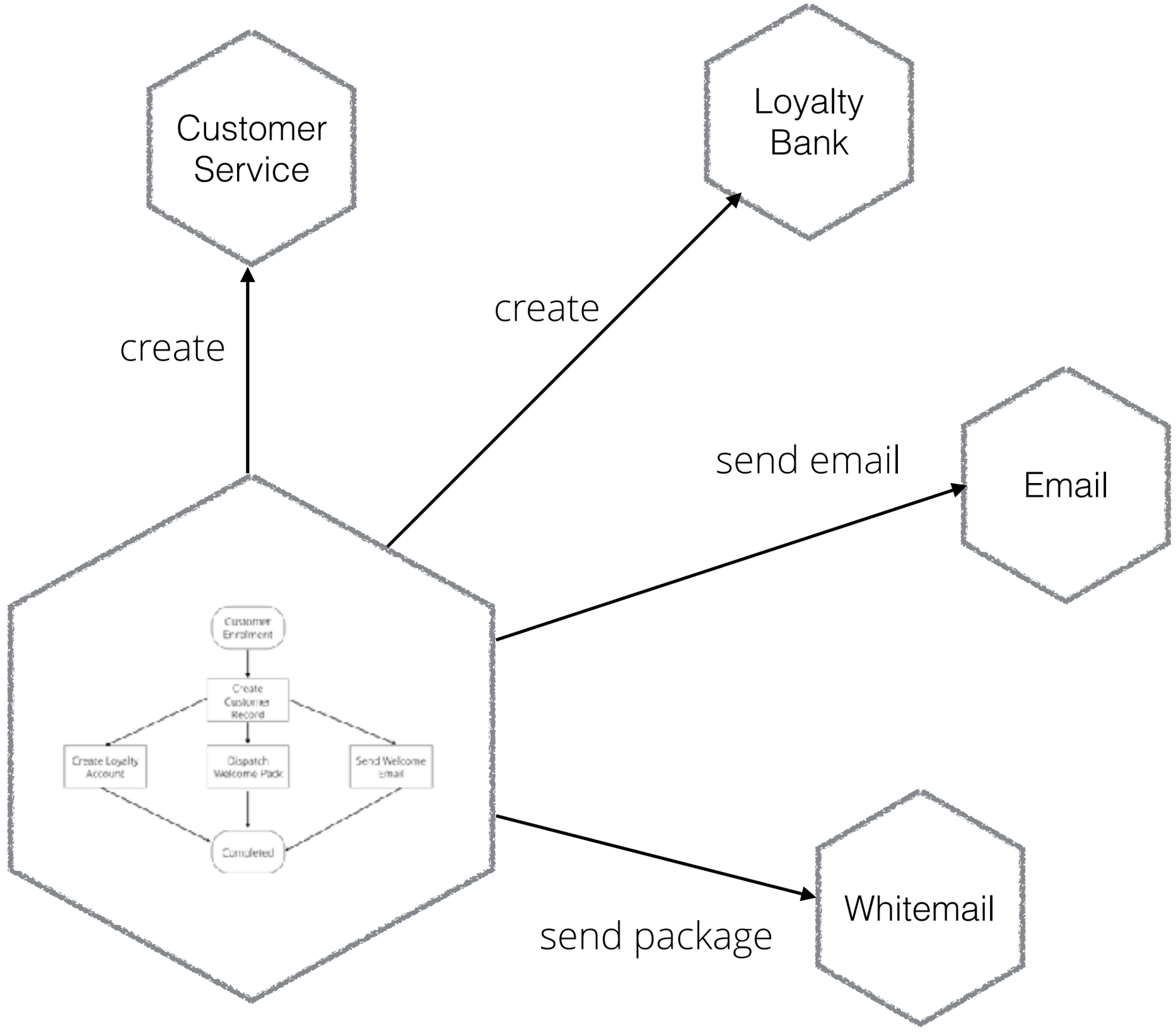
ORCHESTRATION



ORCHESTRATION



ORCHESTRATION



Pros

Pros

Explicit representation of business process

Pros

Explicit representation of business process

Know in-line if there has been a problem

Pros

Explicit representation of business process

Know in-line if there has been a problem

Cons

Pros

Explicit representation of business process

Know in-line if there has been a problem

Cons

Can be fairly coupled

Pros

Explicit representation of business process

Know in-line if there has been a problem

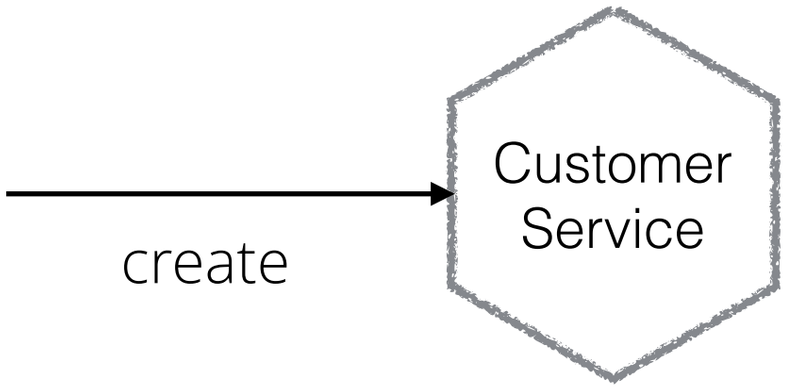
Cons

Can be fairly coupled

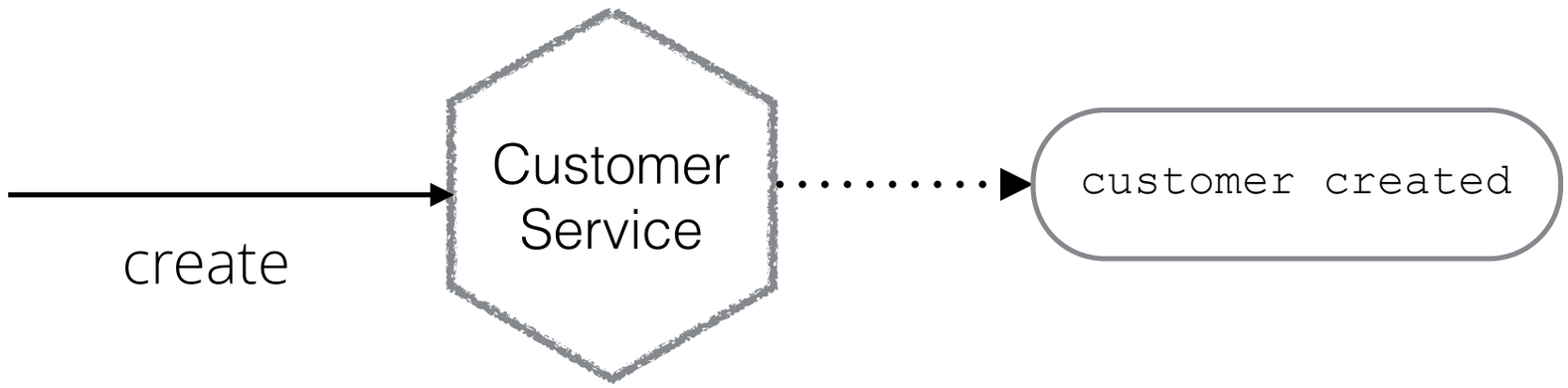
Can lead to overly smart (and dumb services)

CHOREOGRAPHED

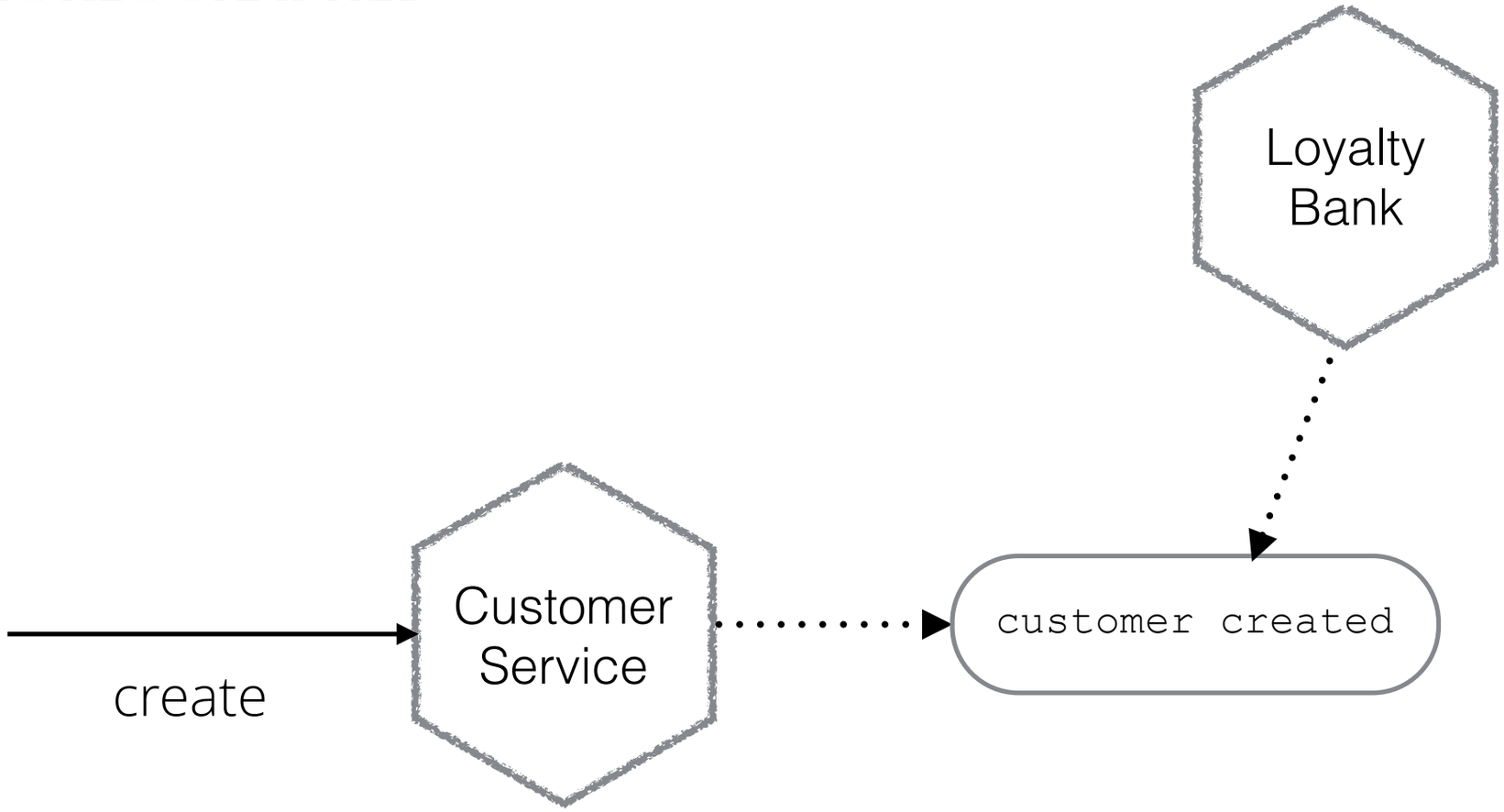
CHOREOGRAPHED



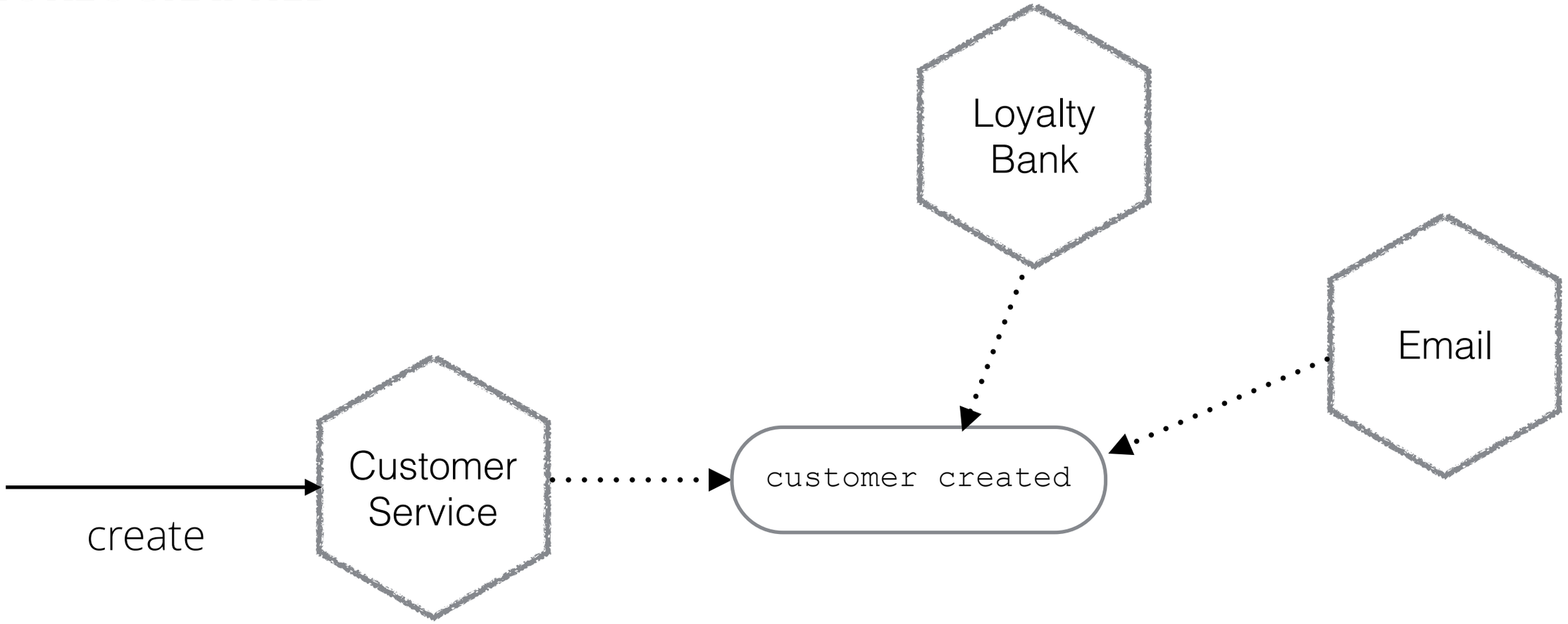
CHOREOGRAPHED



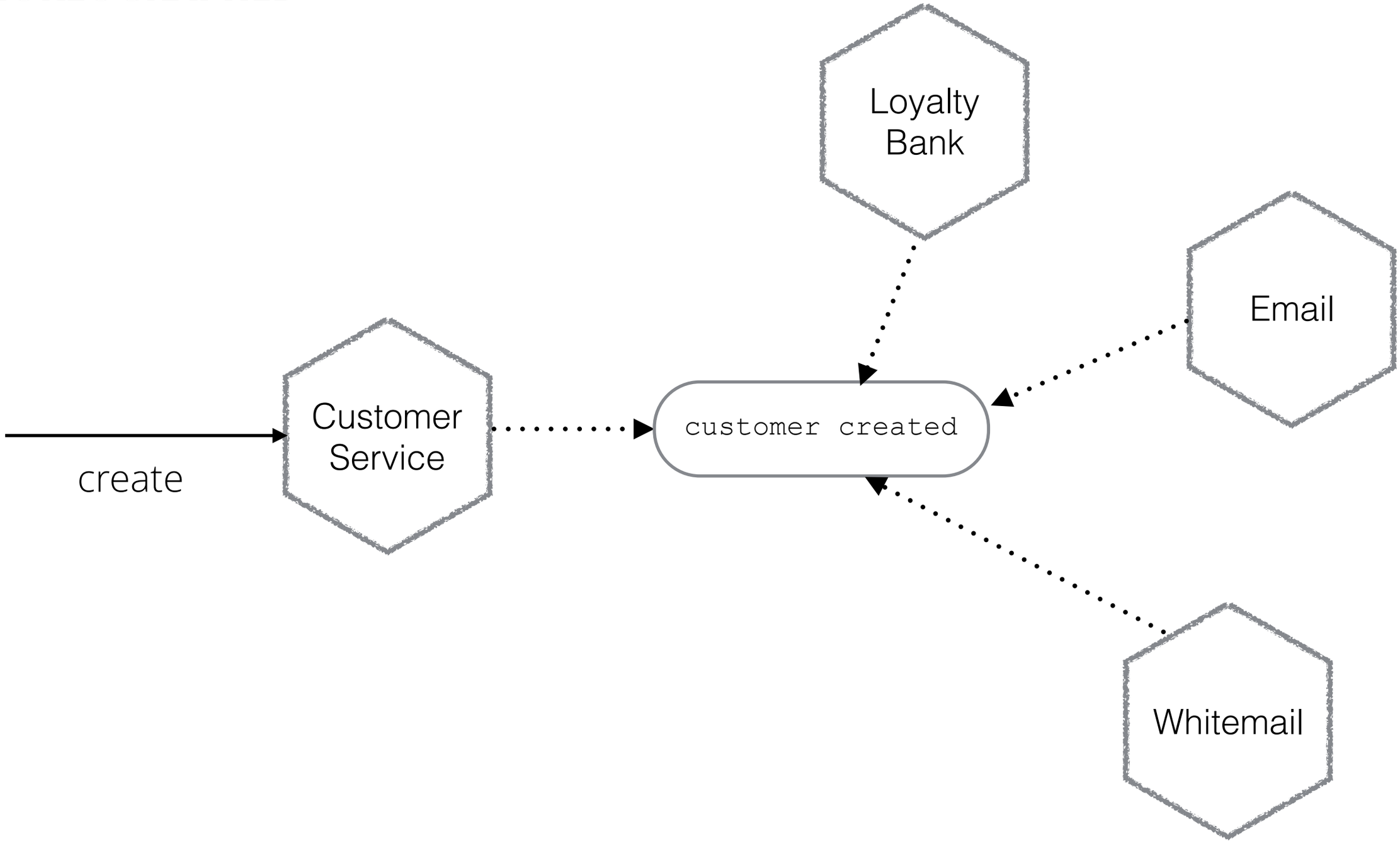
CHOREOGRAPHED



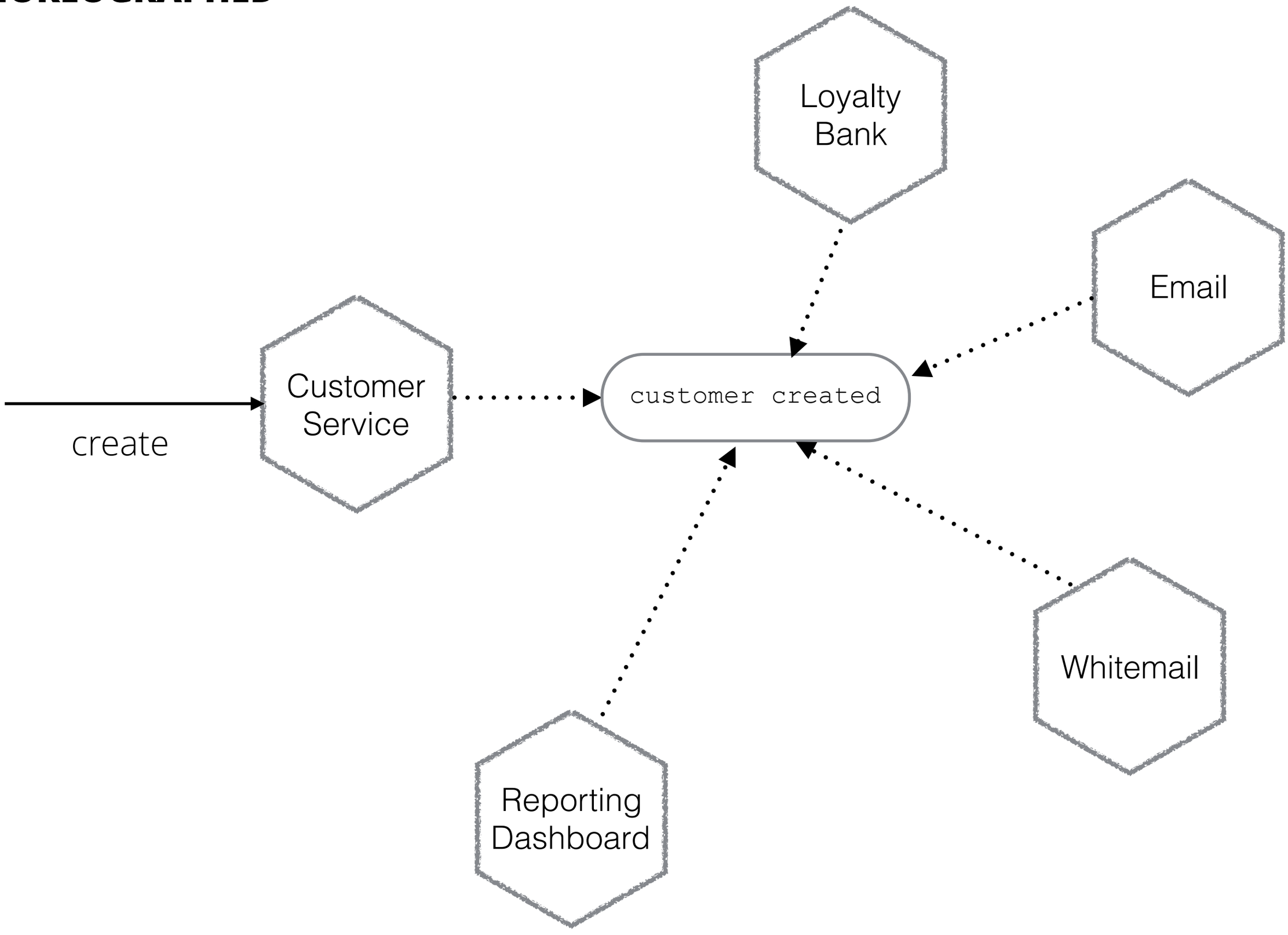
CHOREOGRAPHED



CHOREOGRAPHED



CHOREOGRAPHED



Pros

Pros

Highly decoupled

Pros

Highly decoupled

Evenly distributed smarts

Pros

Highly decoupled

Evenly distributed smarts

Cons

Pros

Highly decoupled

Evenly distributed smarts

Cons

Lost explicit business process mapping

Pros

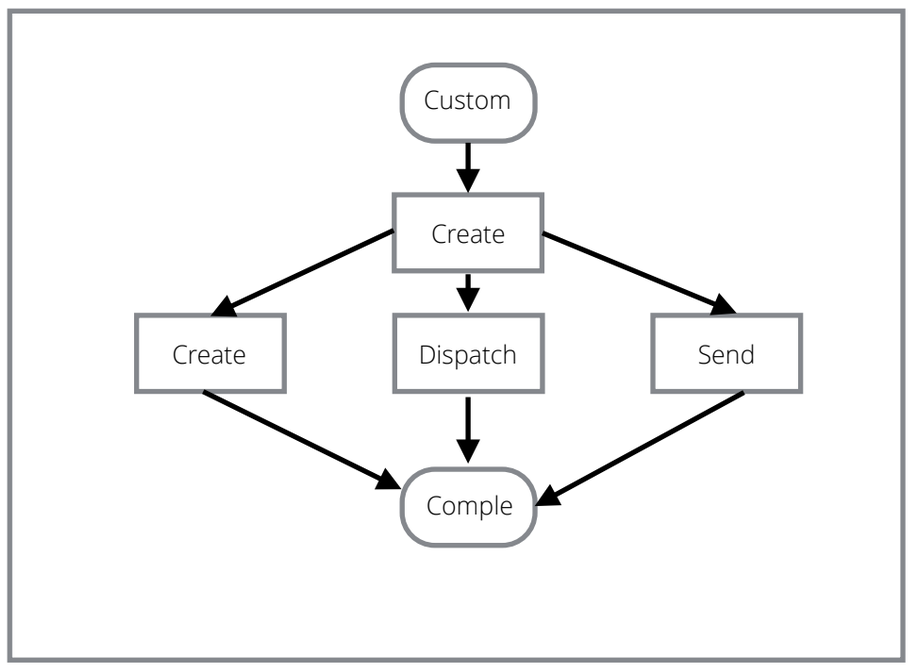
Highly decoupled

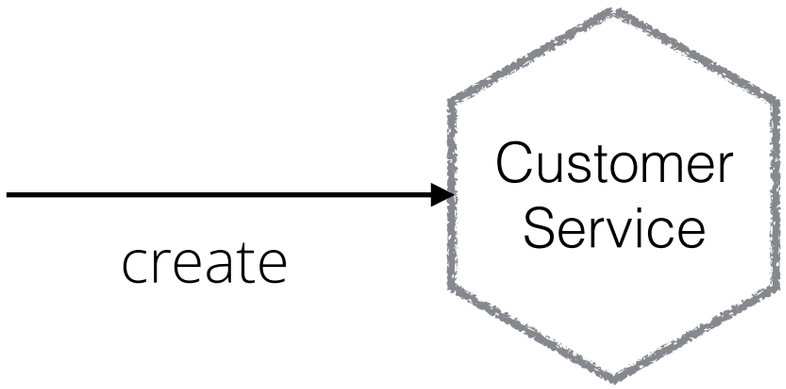
Evenly distributed smarts

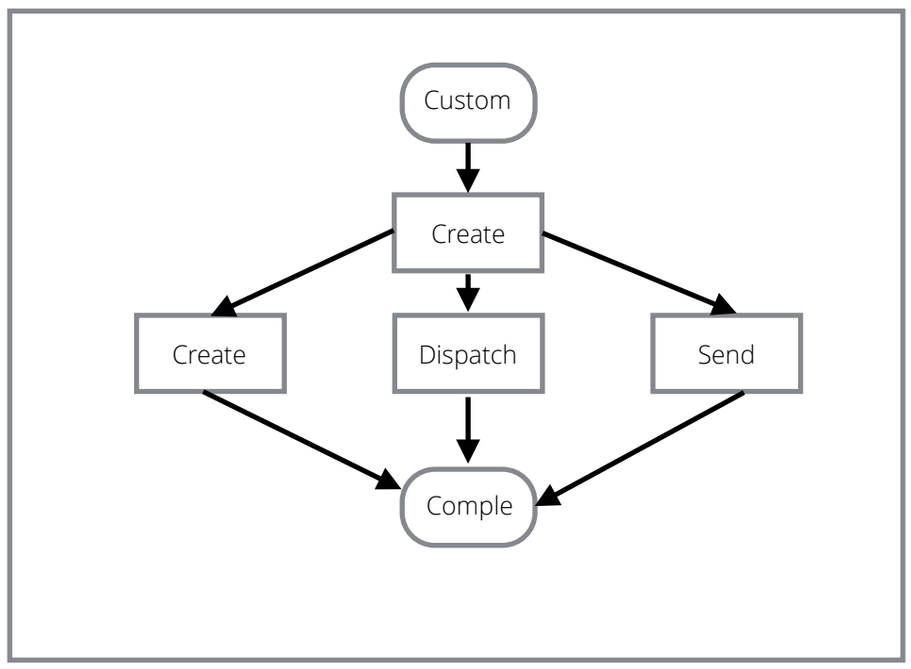
Cons

Lost explicit business process mapping

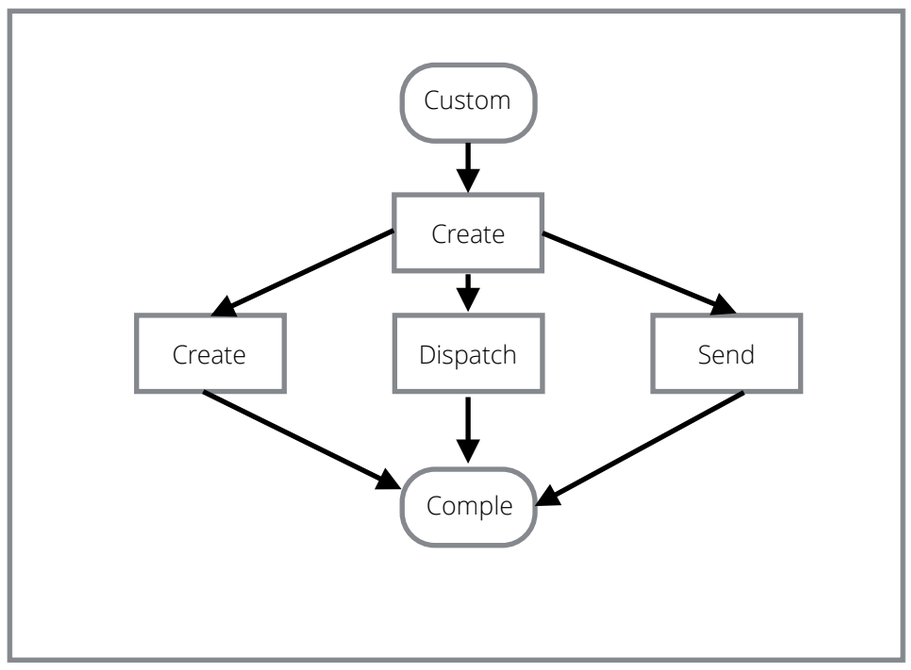
Understanding completion or error states is complex

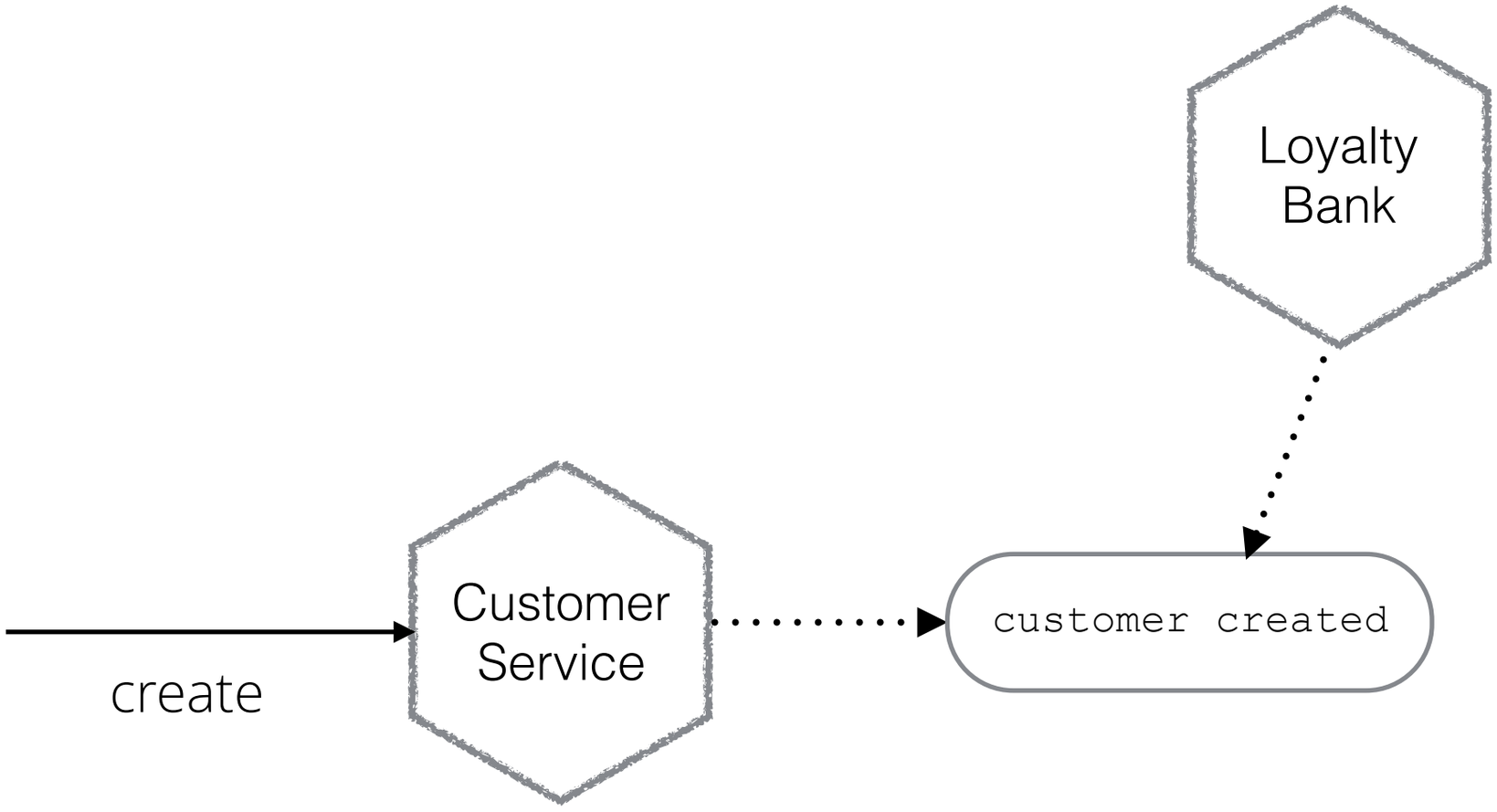

Custom
Create
Create
Dispatch
Send
Comple

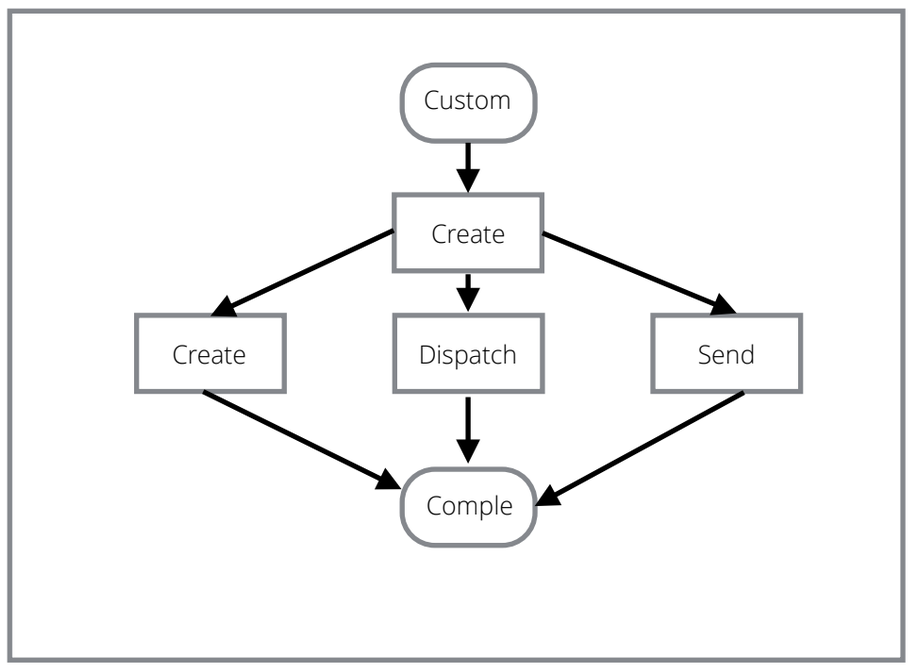

Customer
Service
create

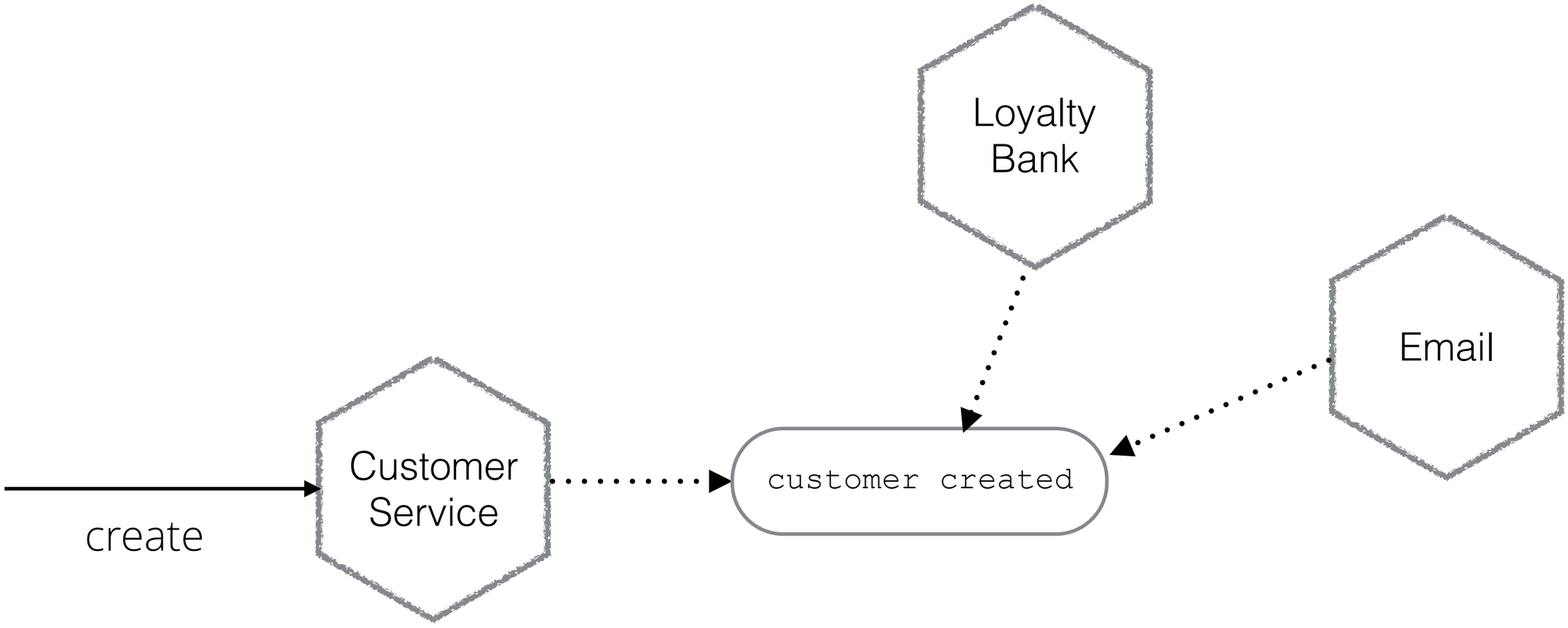

Custom
Create
Create
Dispatch
Send
Comple

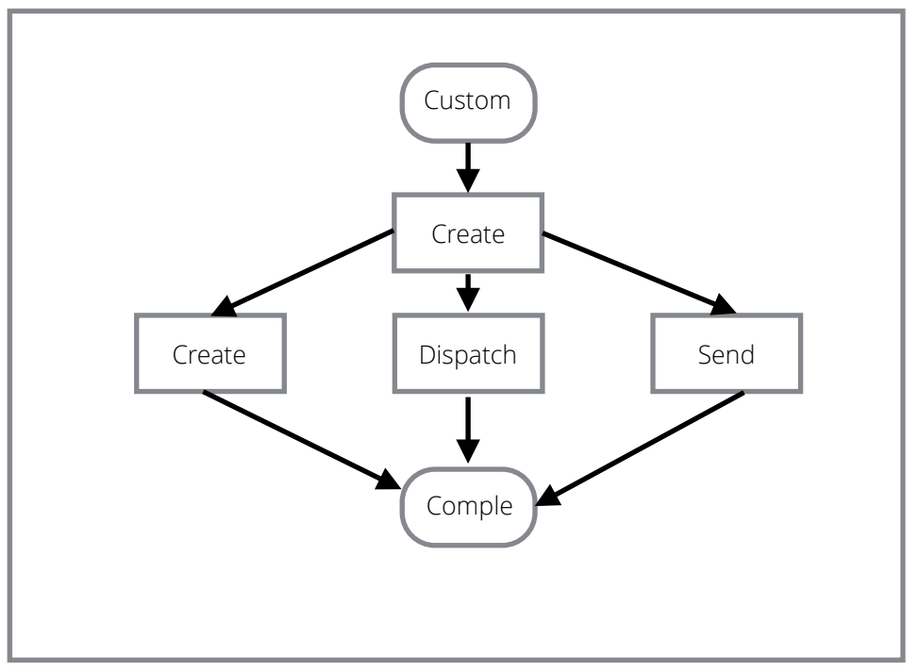

Customer
Service
create
customer created

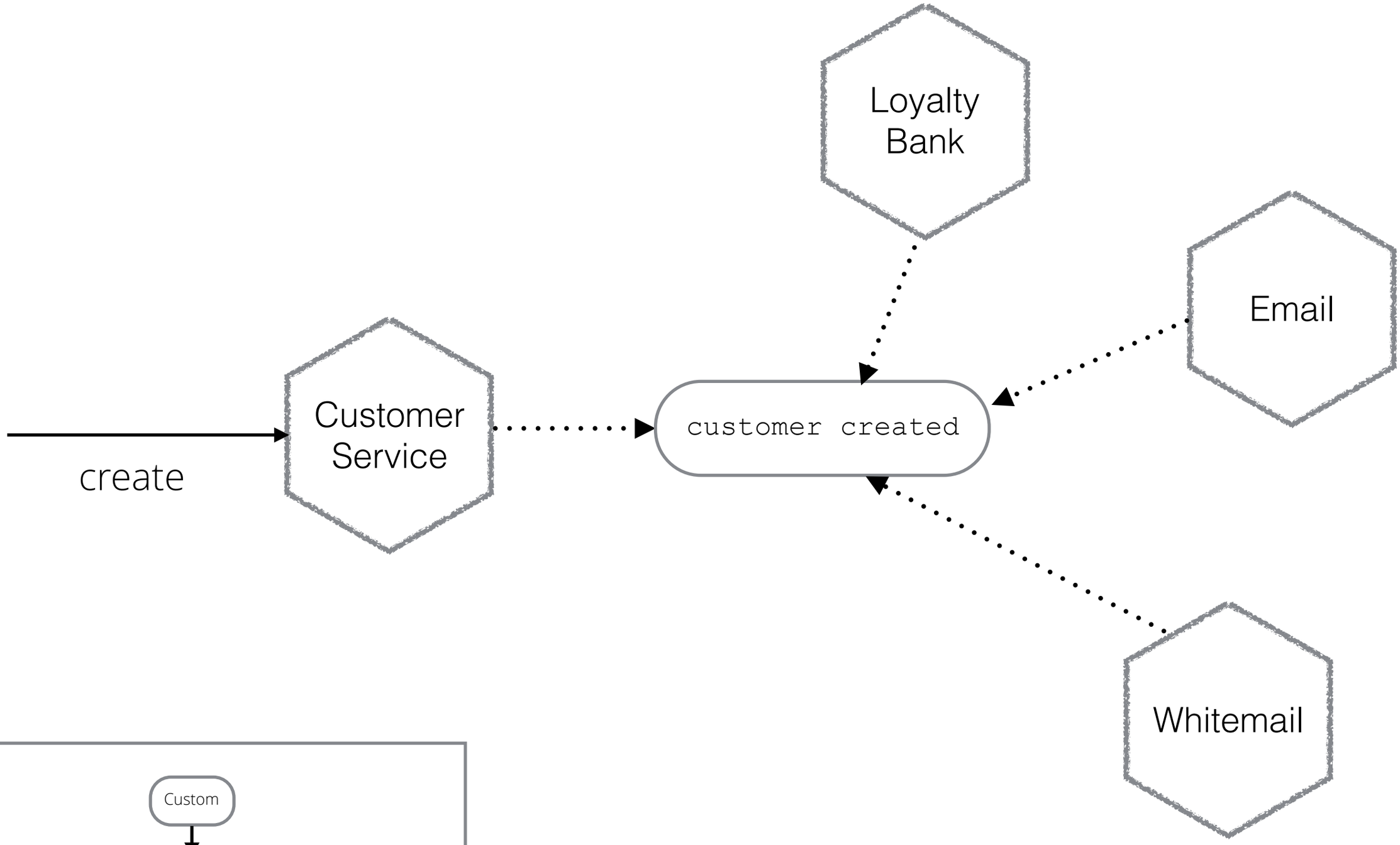

Custom
Create
Create
Dispatch
Send
Comple

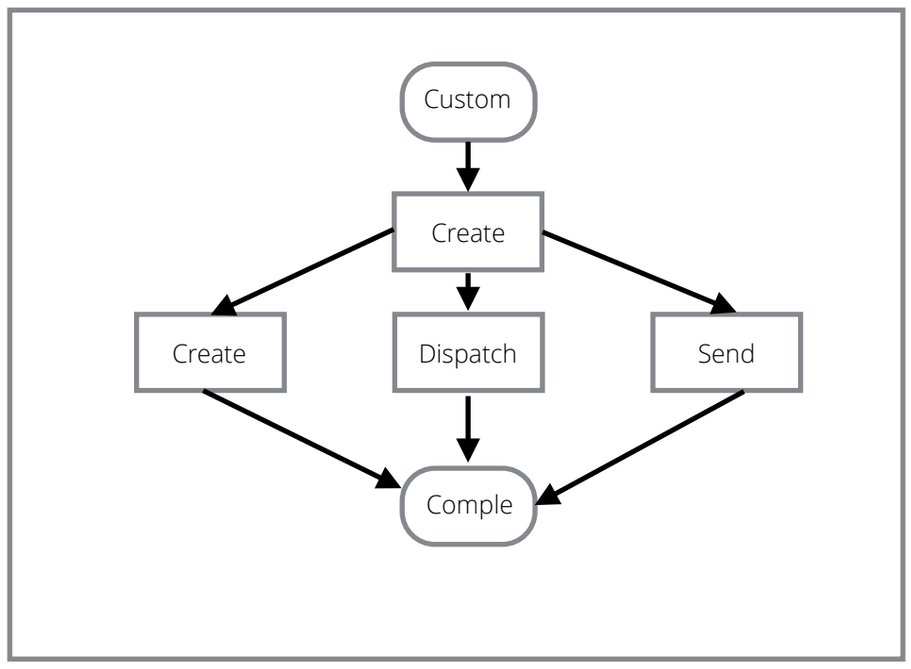

Loyalty
Bank
Customer
Service
customer created
create


Custom
Create
Create
Dispatch
Send
Comple


Loyalty
Bank
Email
Customer
Service
customer created
create


Custom
Create
Create
Dispatch
Send
Comple


Loyalty Bank
Email
Customer Service
customer created
create
Whitemail


Custom
Create
Create
Dispatch
Send
Comple

Summary

Keep it simple

Think about interaction style first, tech second

Orchestration vs Choreography