What are DS?

→ Systems that run on several nodes and are characterised by partial failure

Request

M_1 M_2 Process
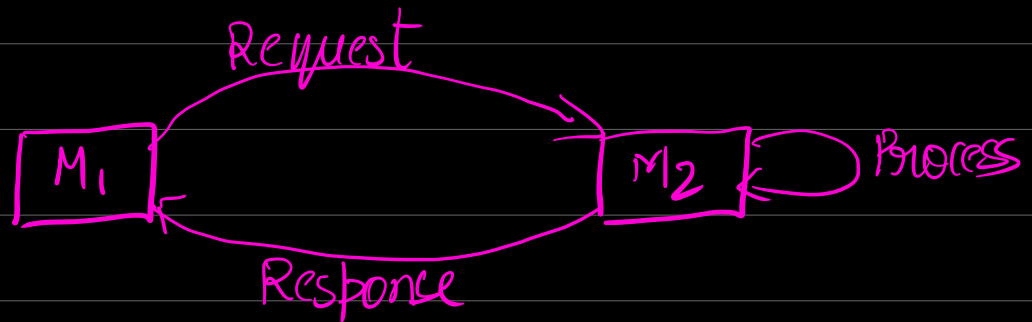
Response

Reasons why M_1 might not receive response

→ Request takes too long

→ Request lost

→ M_2 down

→ M_2 down immediately after reporting it is up

→ M_2 refuses request

→ M_2 response too slow

→ M_2 response lost

Time outs

M_1 ← Max delay time = d — M_2 Max processing time = q

Ideal timeout: $\dfrac{2d}{\downarrow} + \dfrac{n}{\llcorner\to}$ Processing

Round trip

In practice, you will never get the real value for timeout. Values for d & n are estimated using statistics

Why DS even after so many problems?

$\to$ Too much data to fit on single machine
$\to$ Need faster response time $\Rightarrow$ more processing power
$\to$ Scalability
$\to$ Fault tolerence (redundancy)