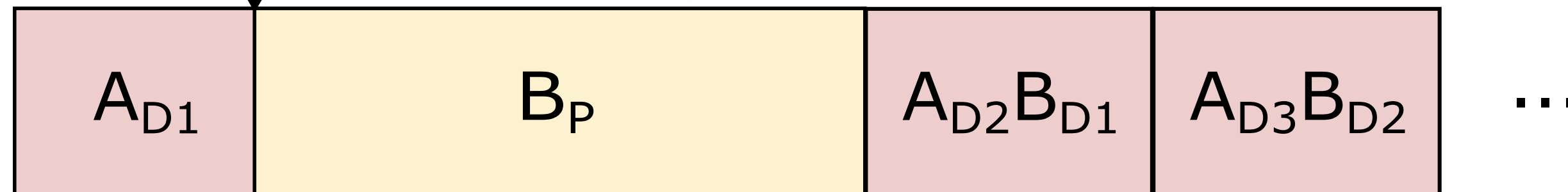


Long Request B
Arrives

Request A Blocked

**Standard Continuous
Batching**



Chunked Prefills
Unblock Request A

**Batching with
Chunked Prefill**

