



# Pipeline and Hybrid Parallelism

Abhinav Bhatele, Daniel Nichols



UNIVERSITY OF  
MARYLAND

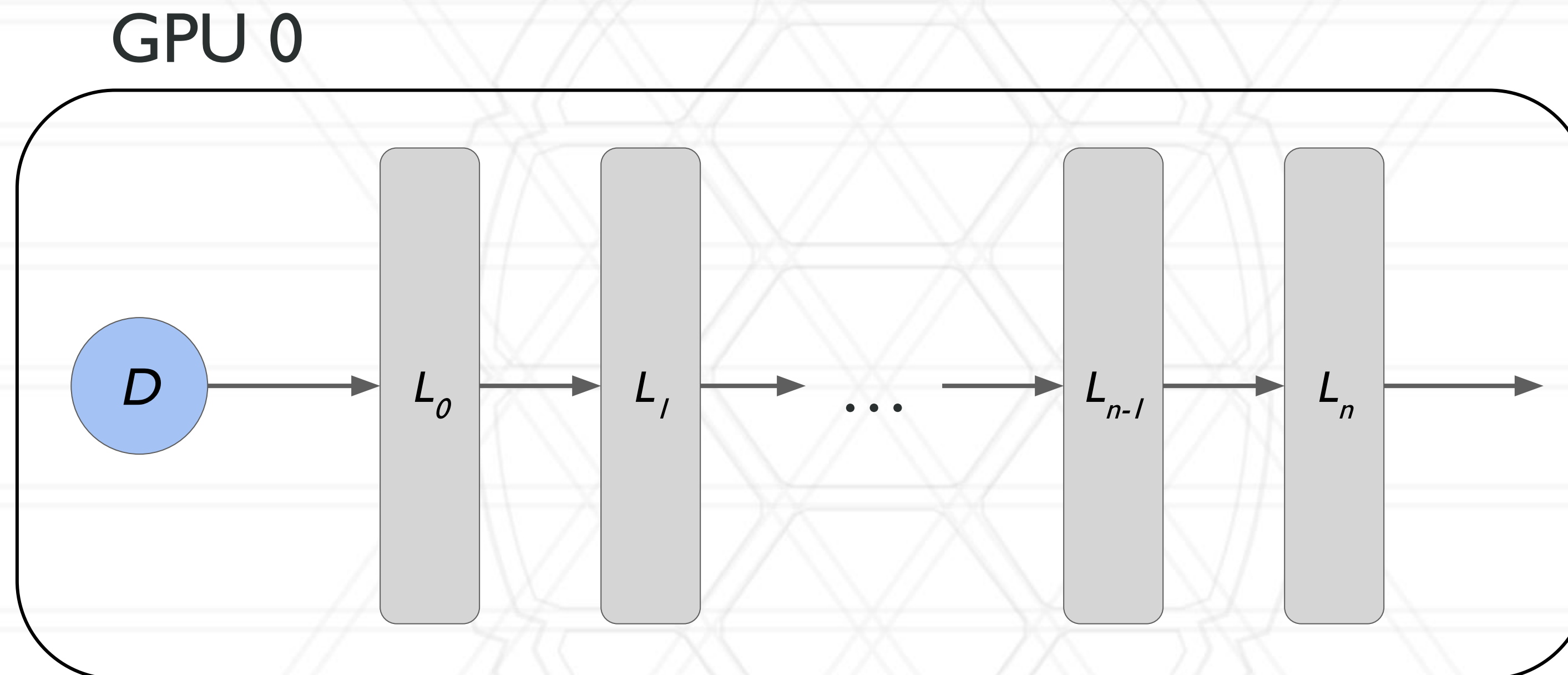
# Announcements

---

- Project proposals and groups due Friday, March 7<sup>th</sup>
- Assignment 2 due March 14<sup>th</sup> (with extension to March 17<sup>th</sup>)

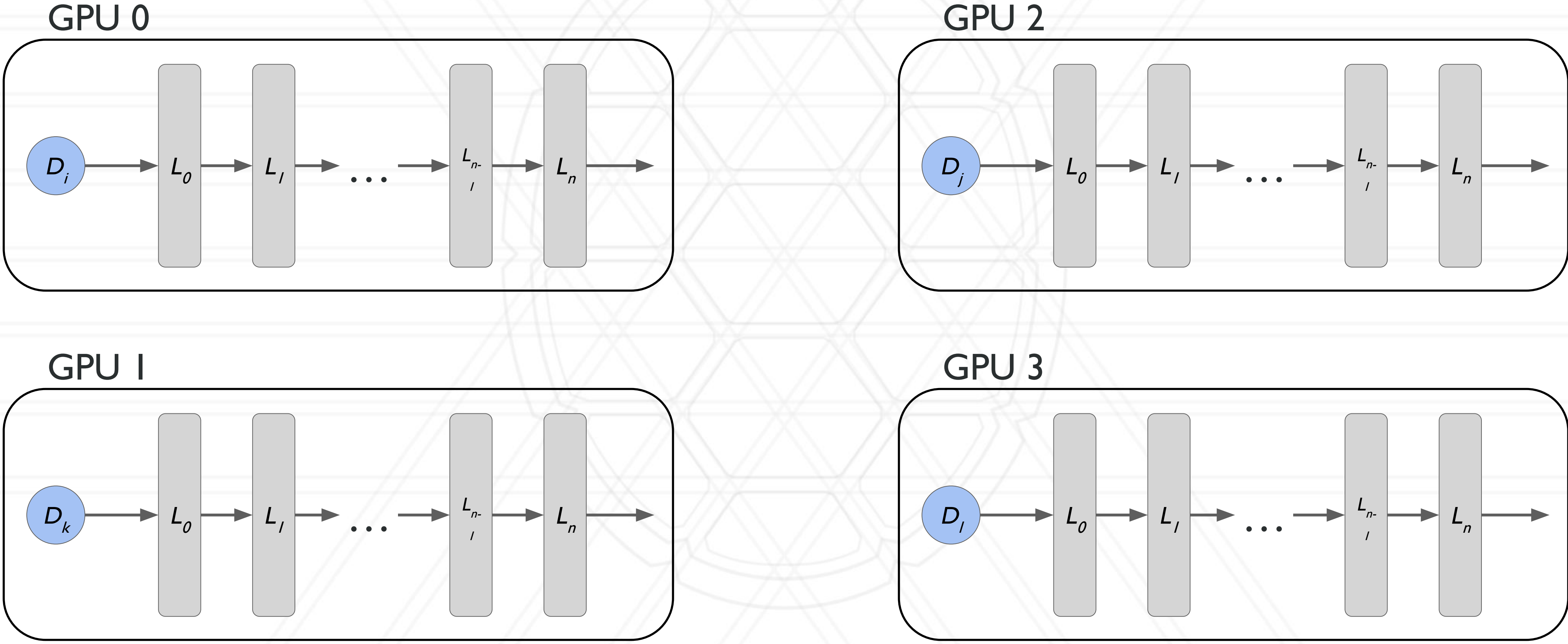
# Sequential Training

---

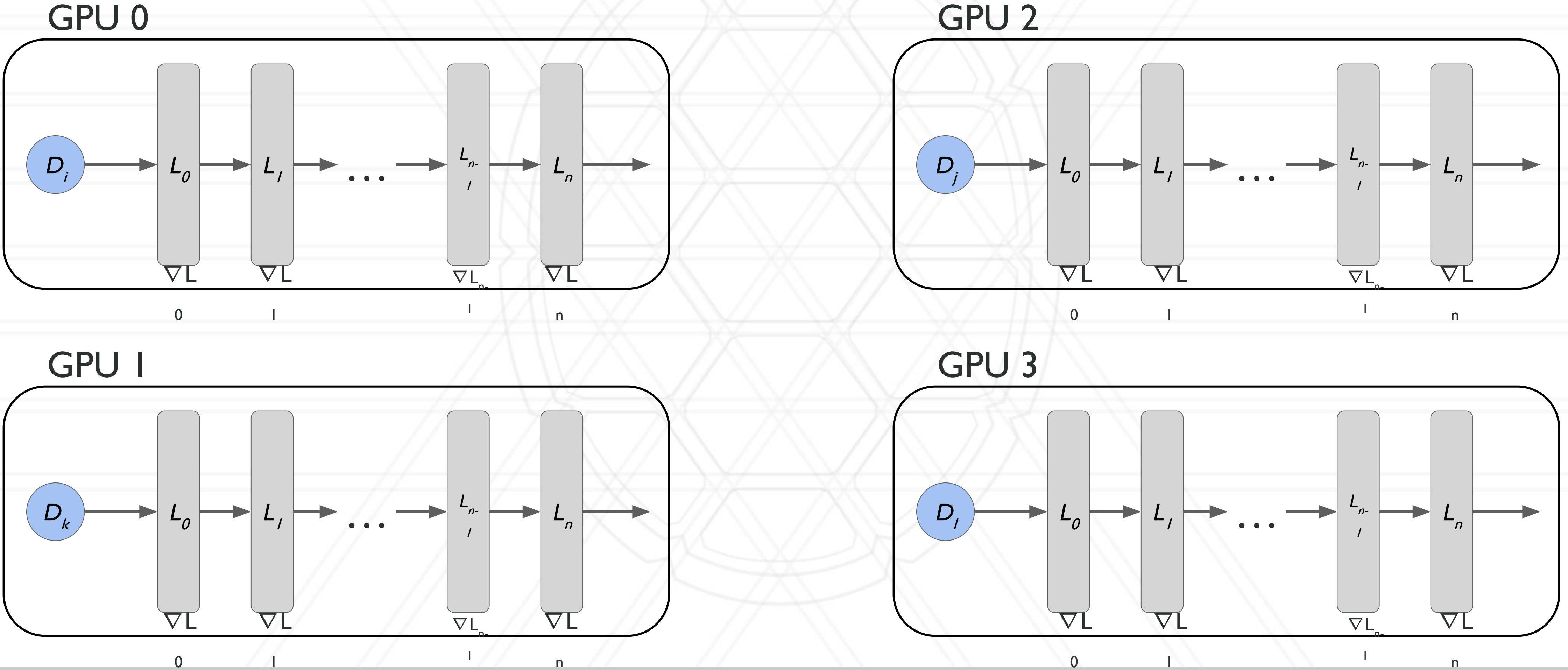




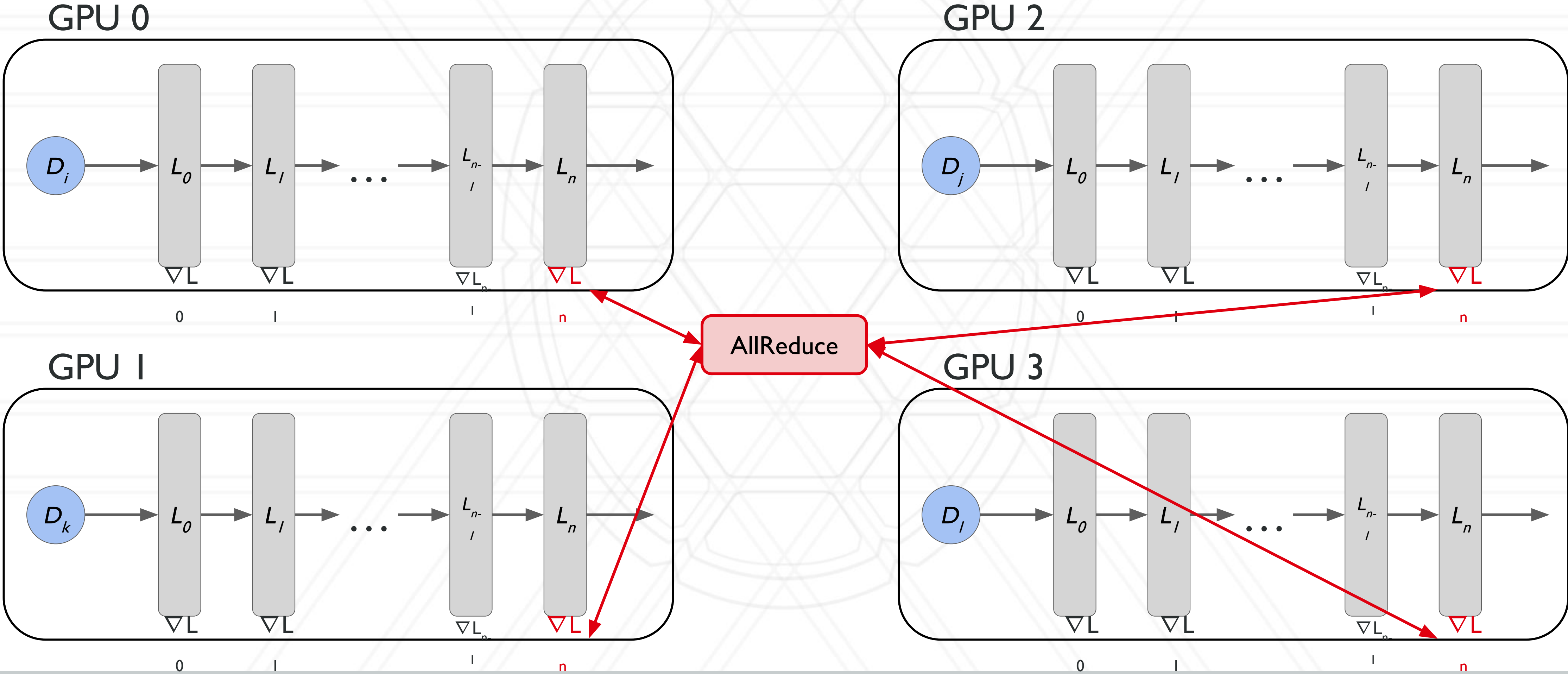
# Data Parallel Training



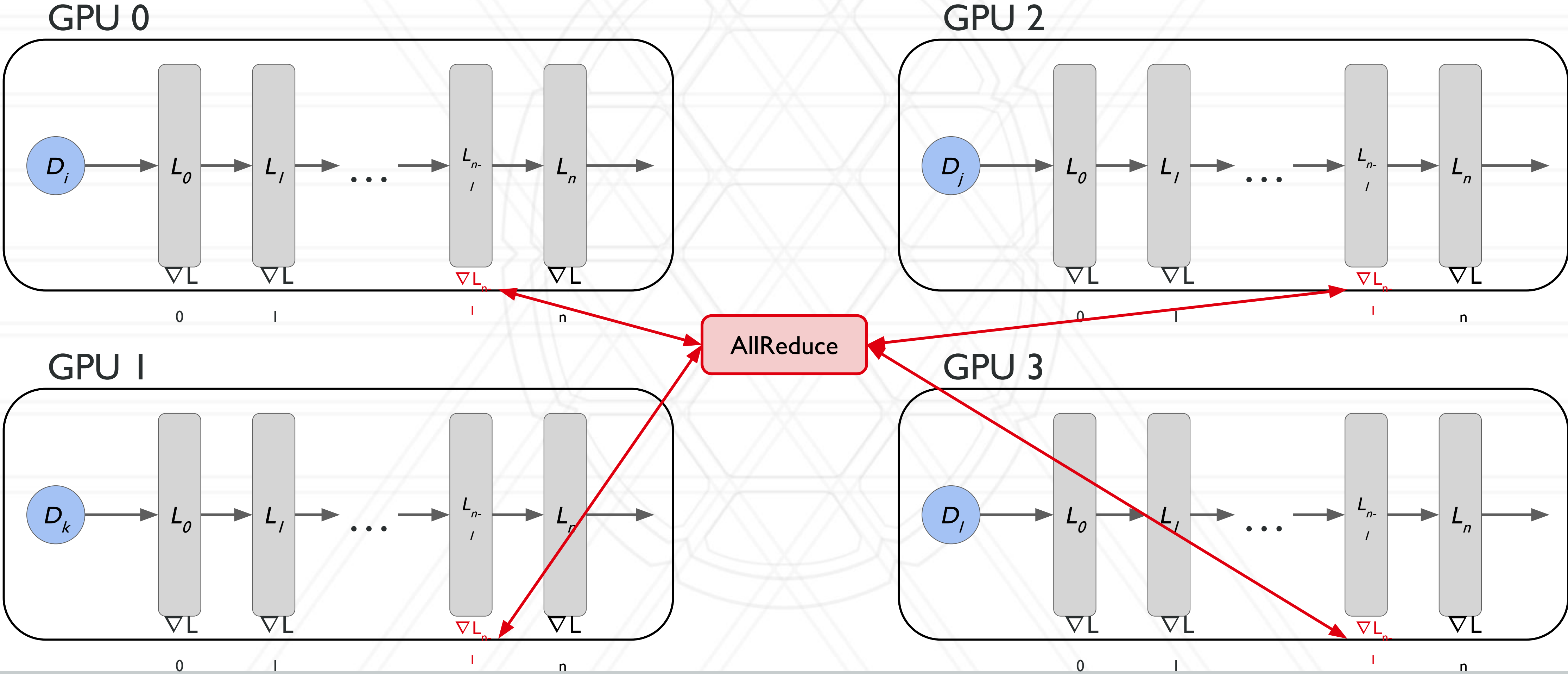
# Data Parallel Training



# Data Parallel Training

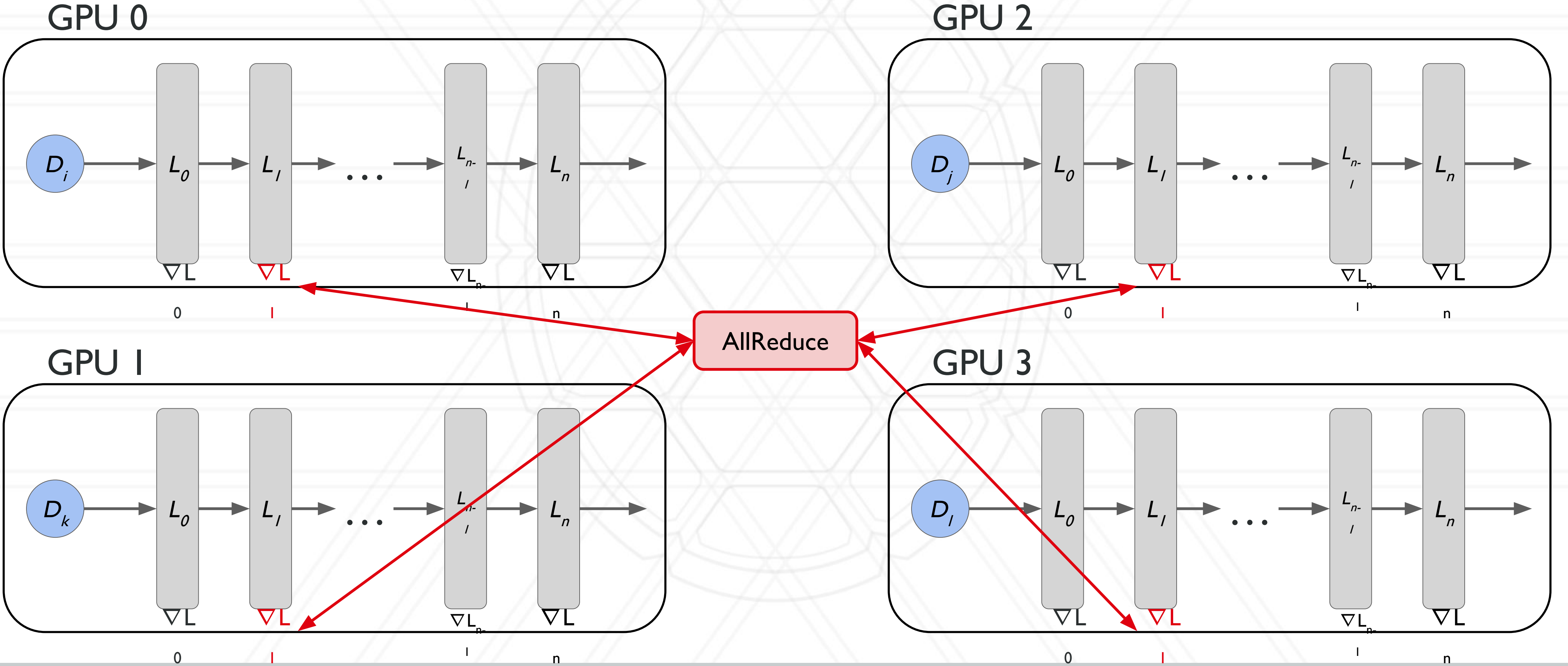


# Data Parallel Training



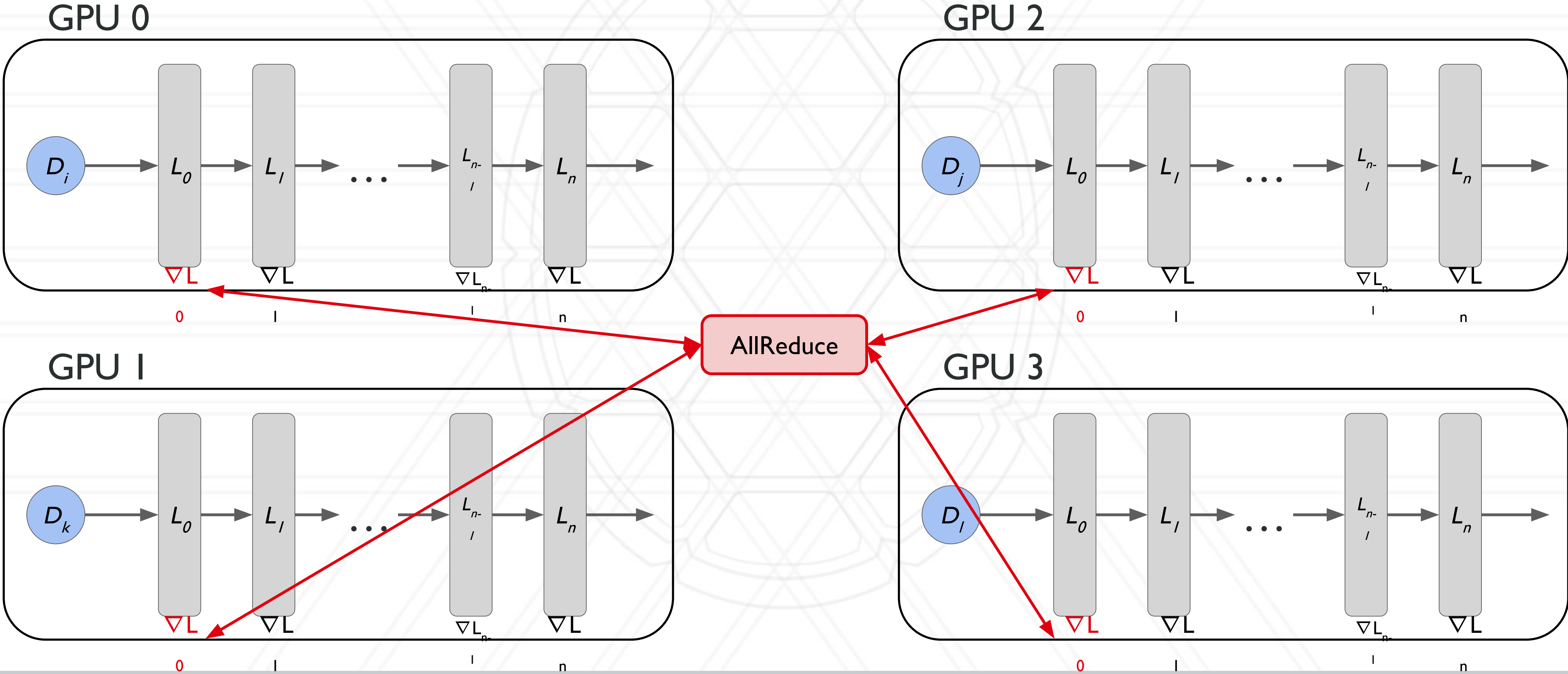


# Data Parallel Training



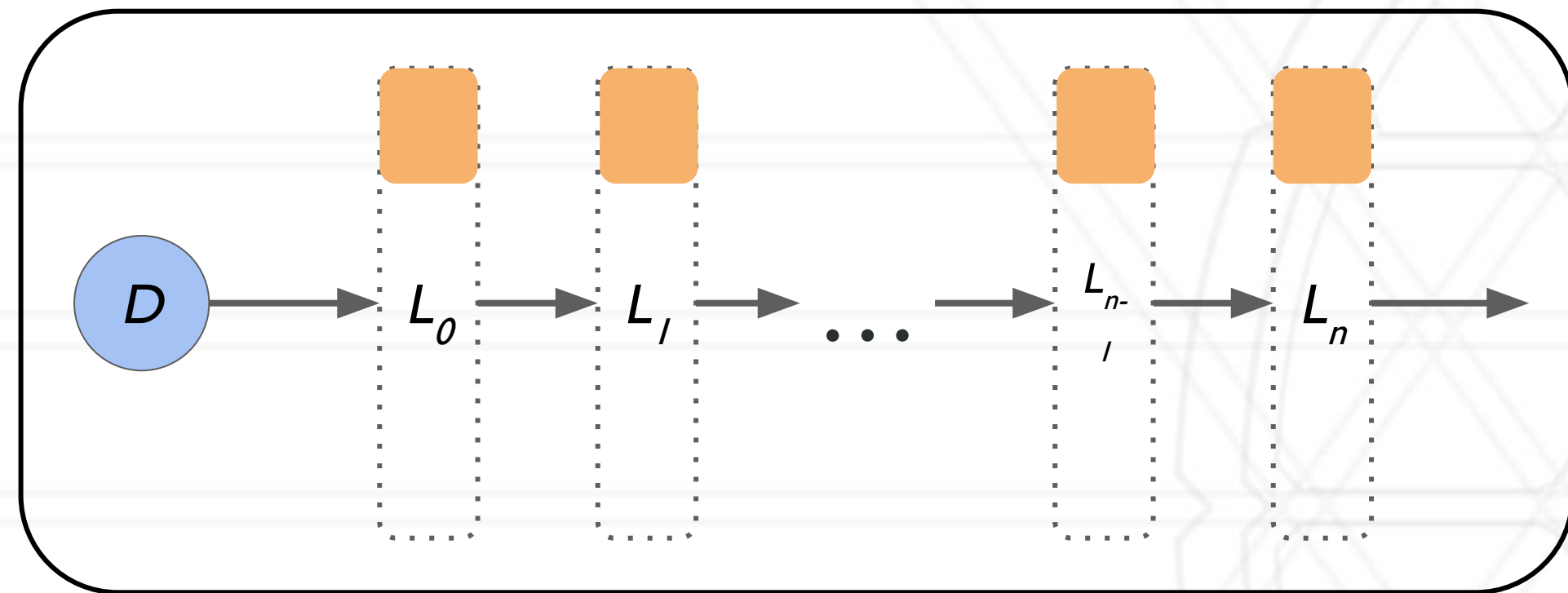


# Data Parallel Training

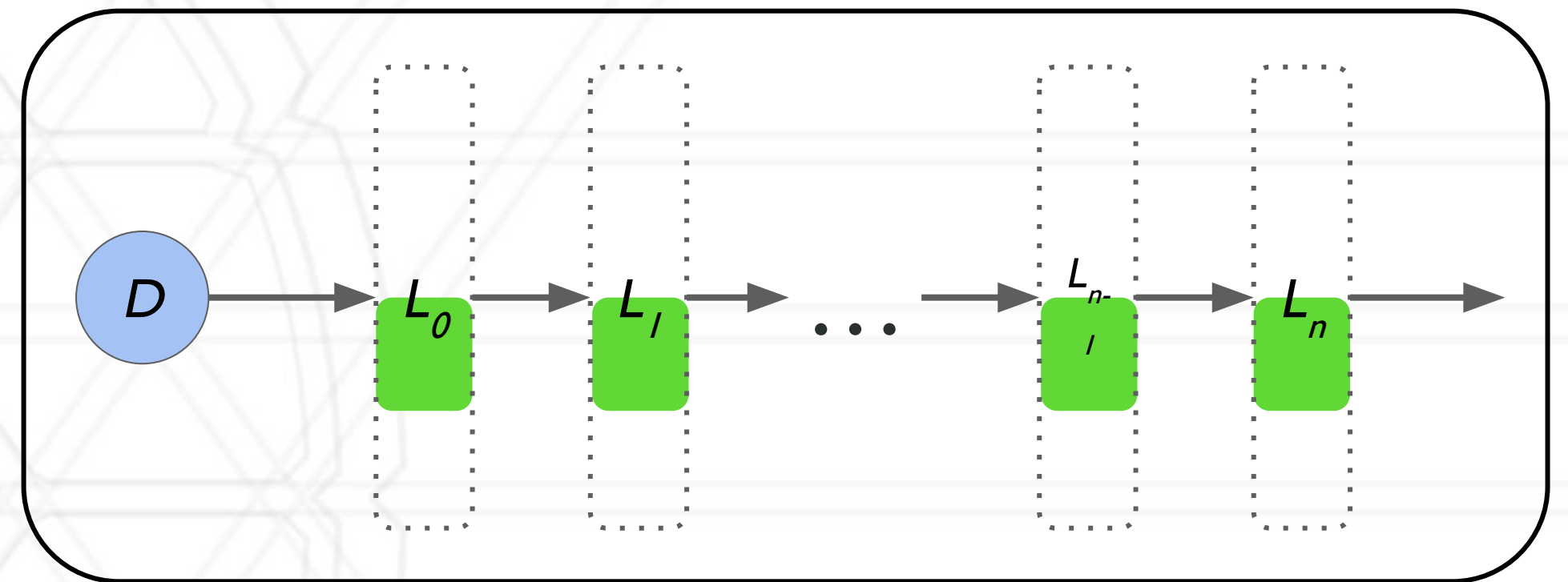


# Tensor Parallel Training

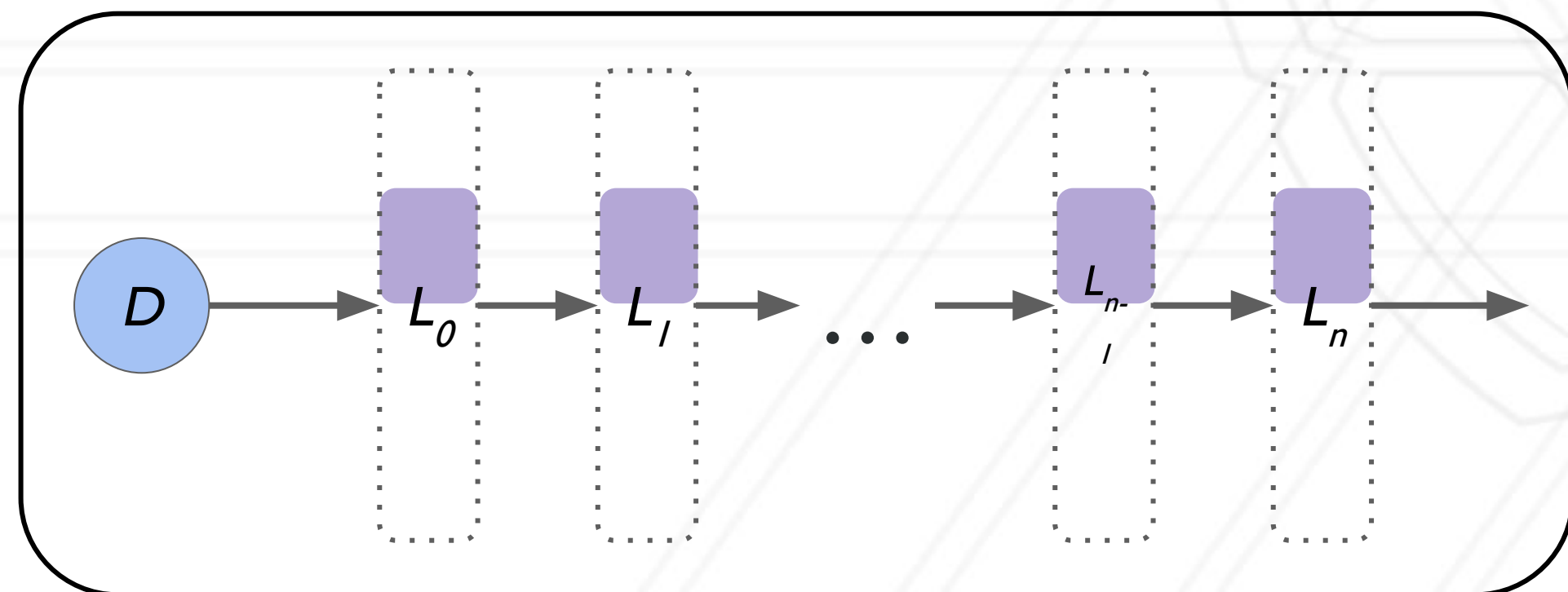
GPU 0



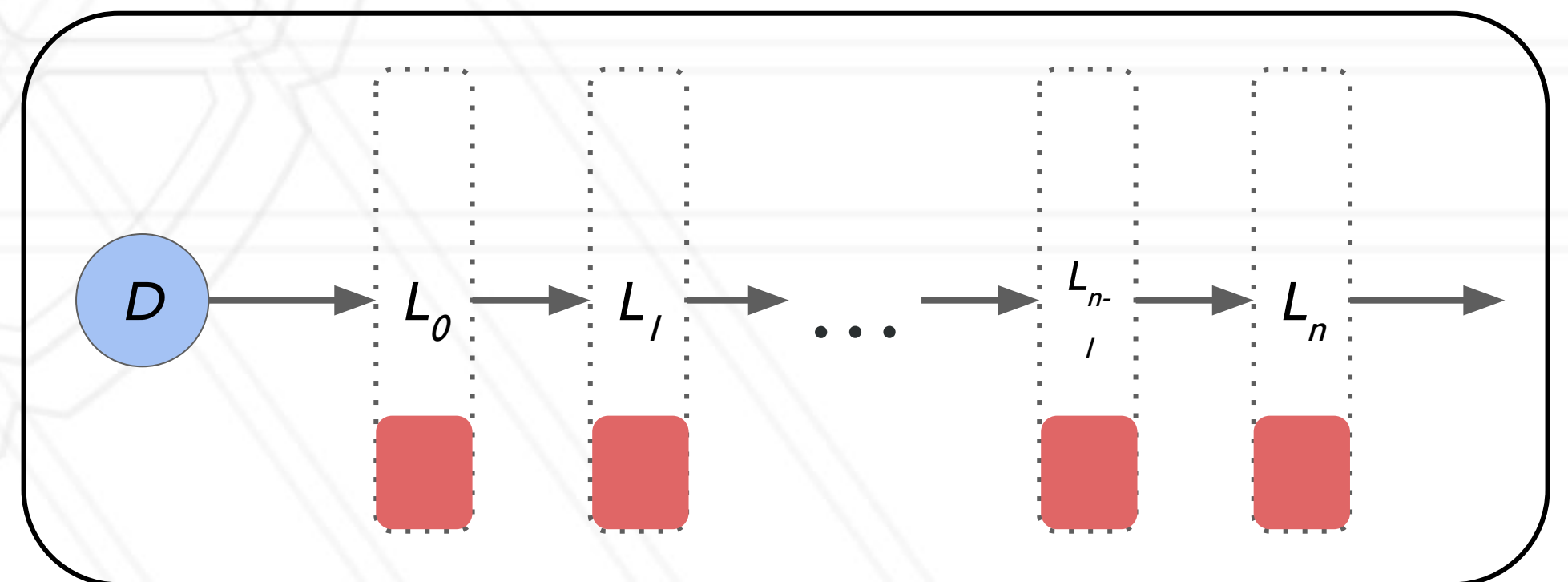
GPU 2



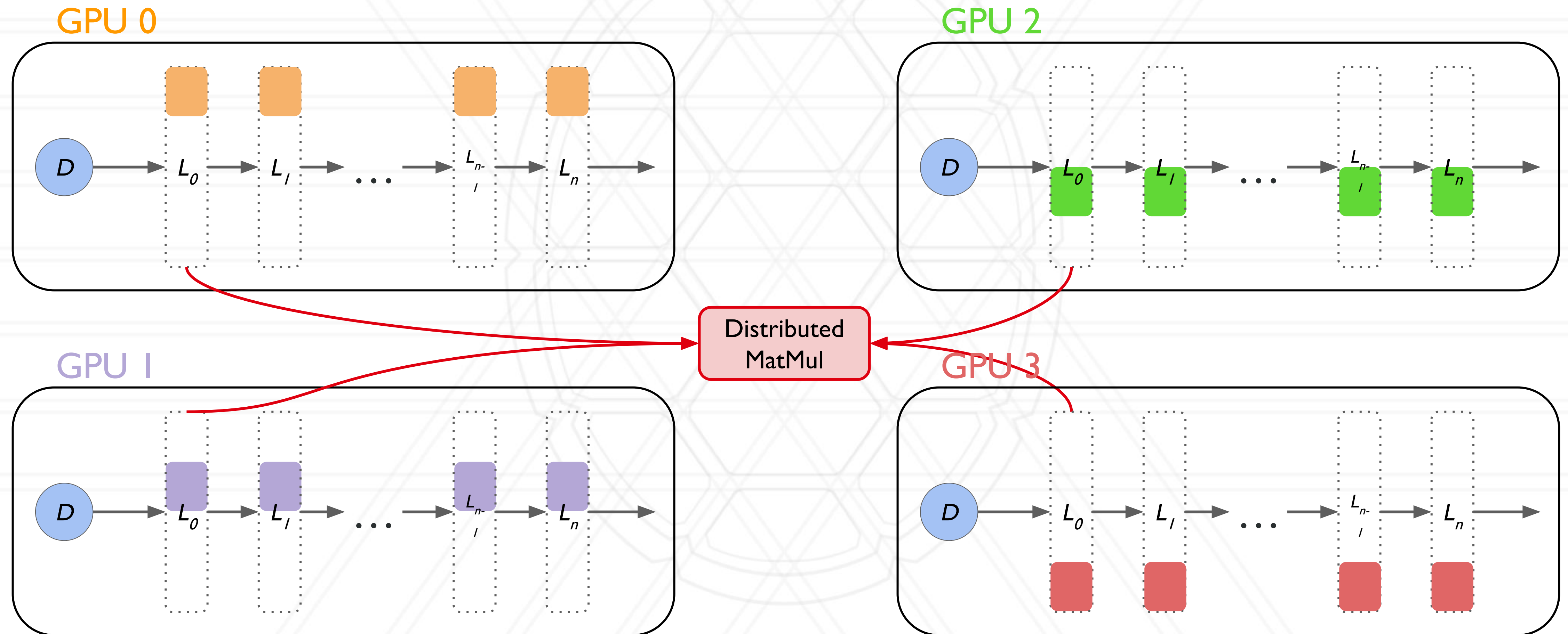
GPU 1



GPU 3

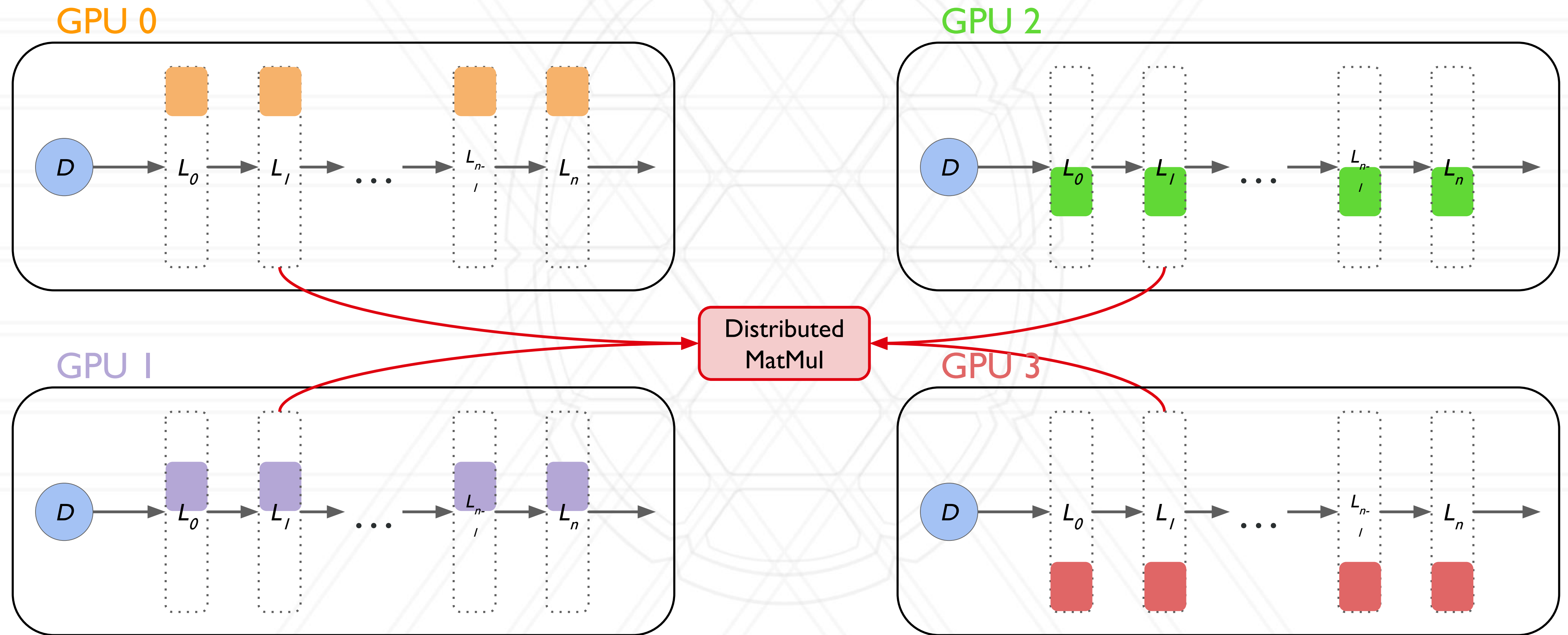


# Tensor Parallel Training

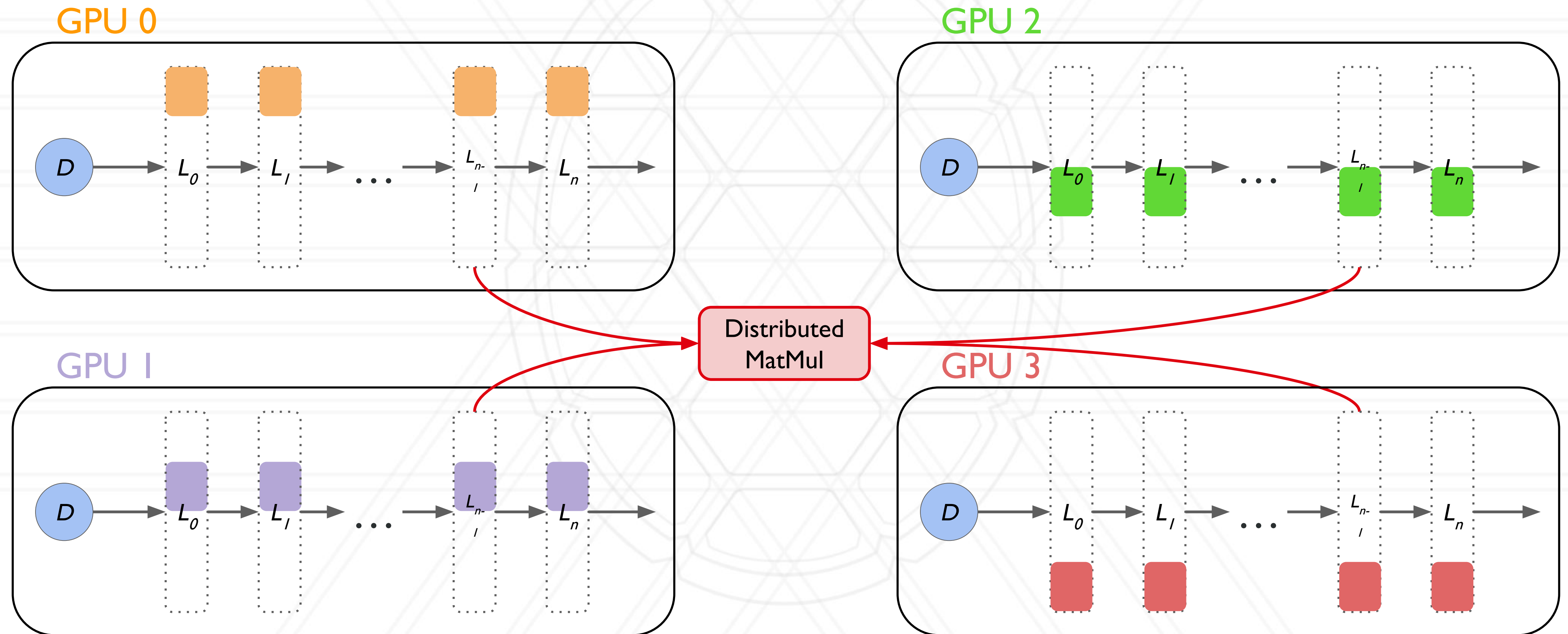




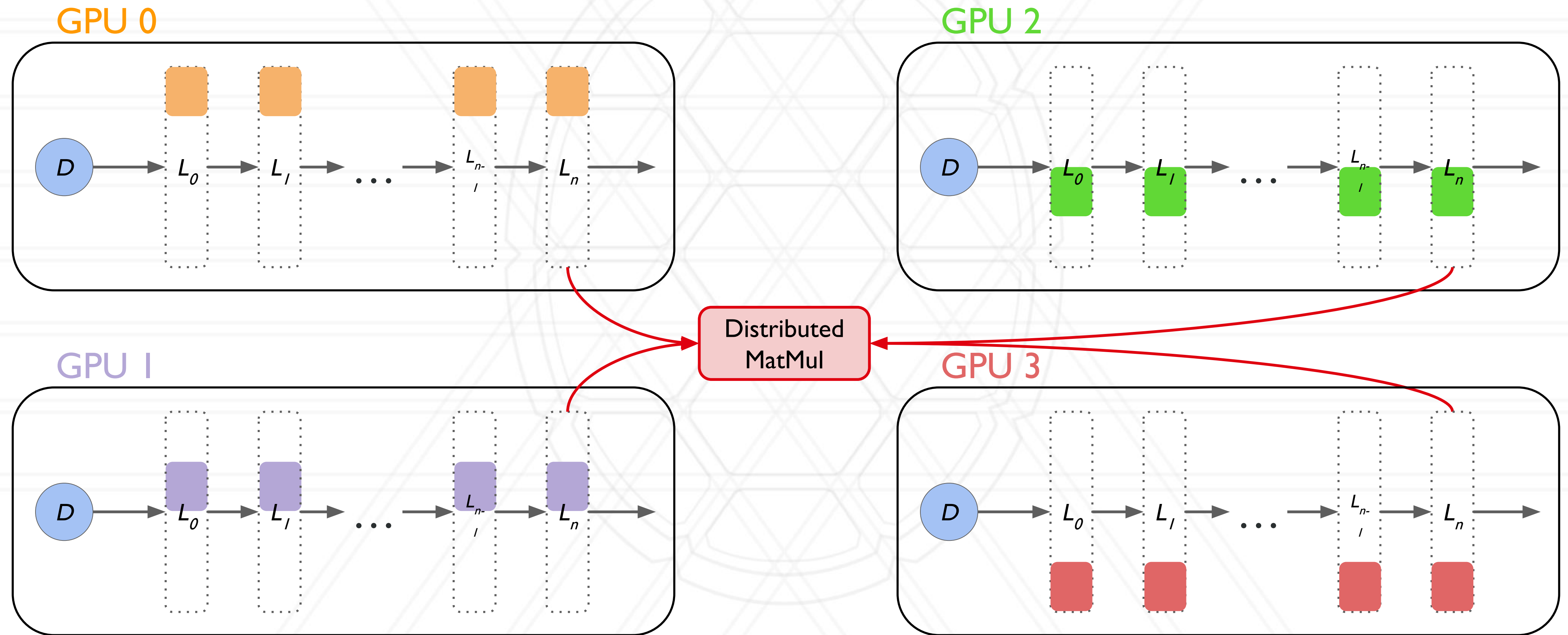
# Tensor Parallel Training



# Tensor Parallel Training



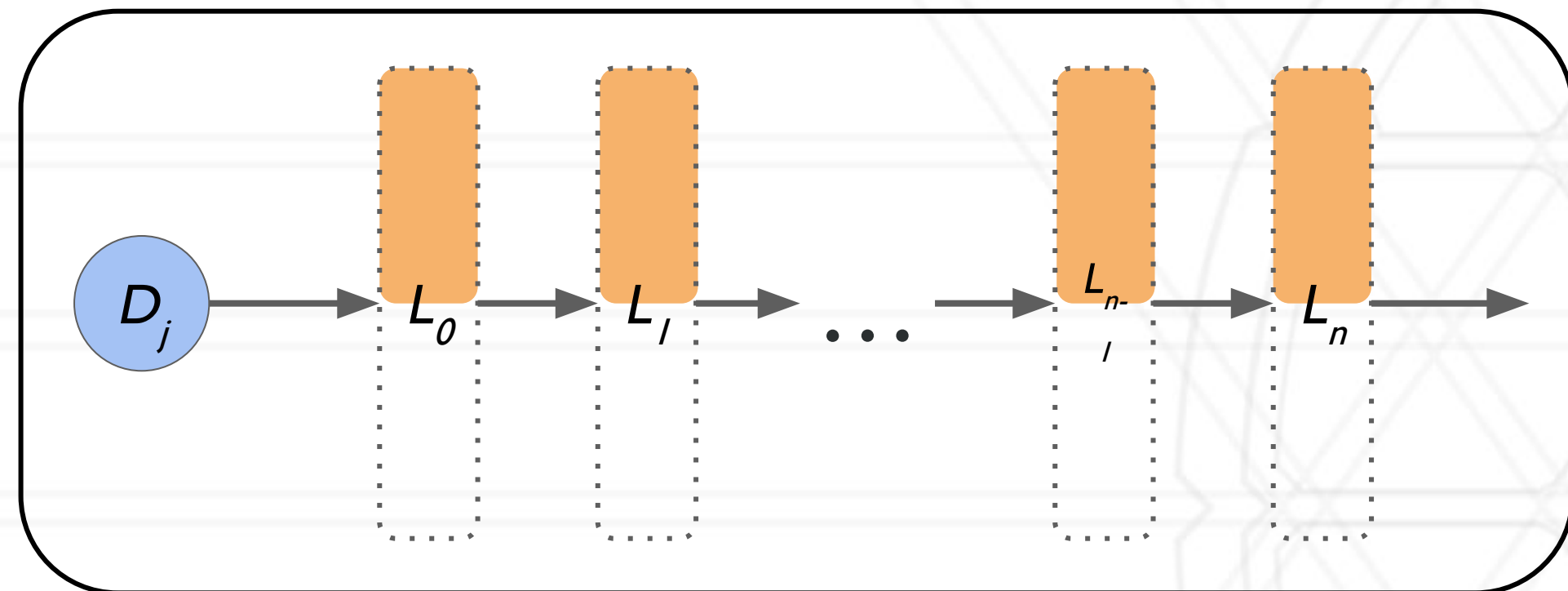
# Tensor Parallel Training



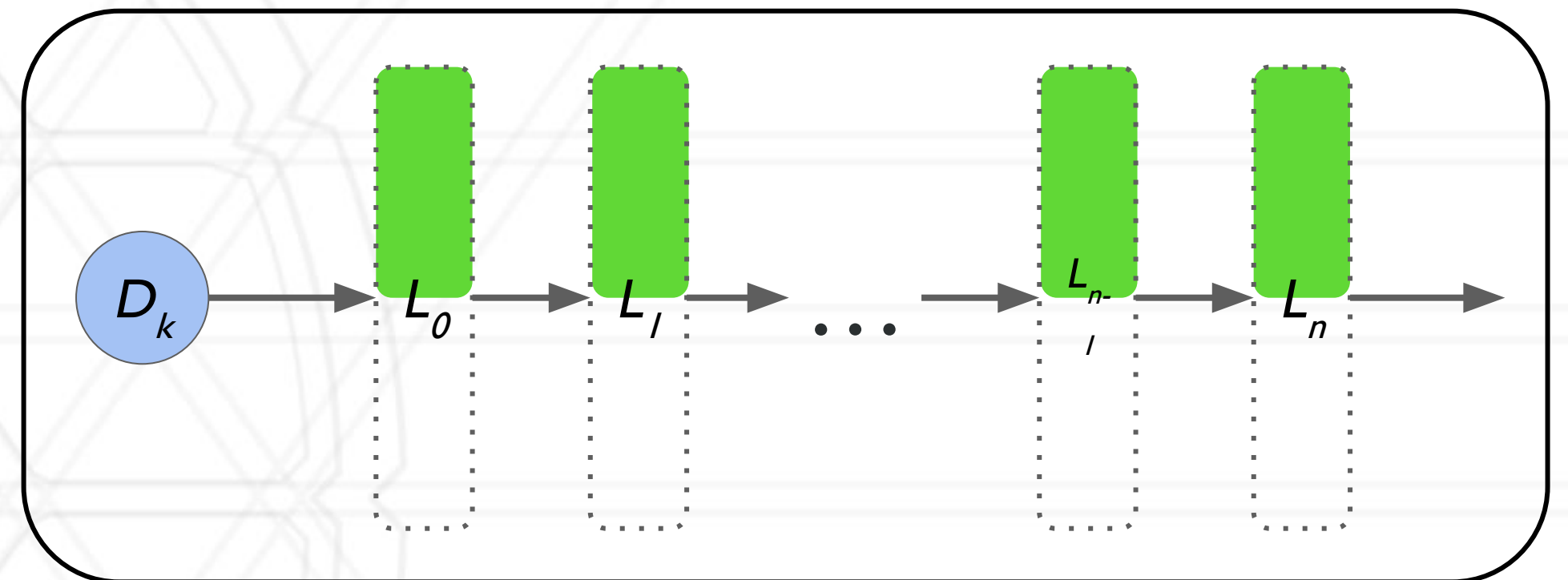


# Tensor + Data Parallel Training

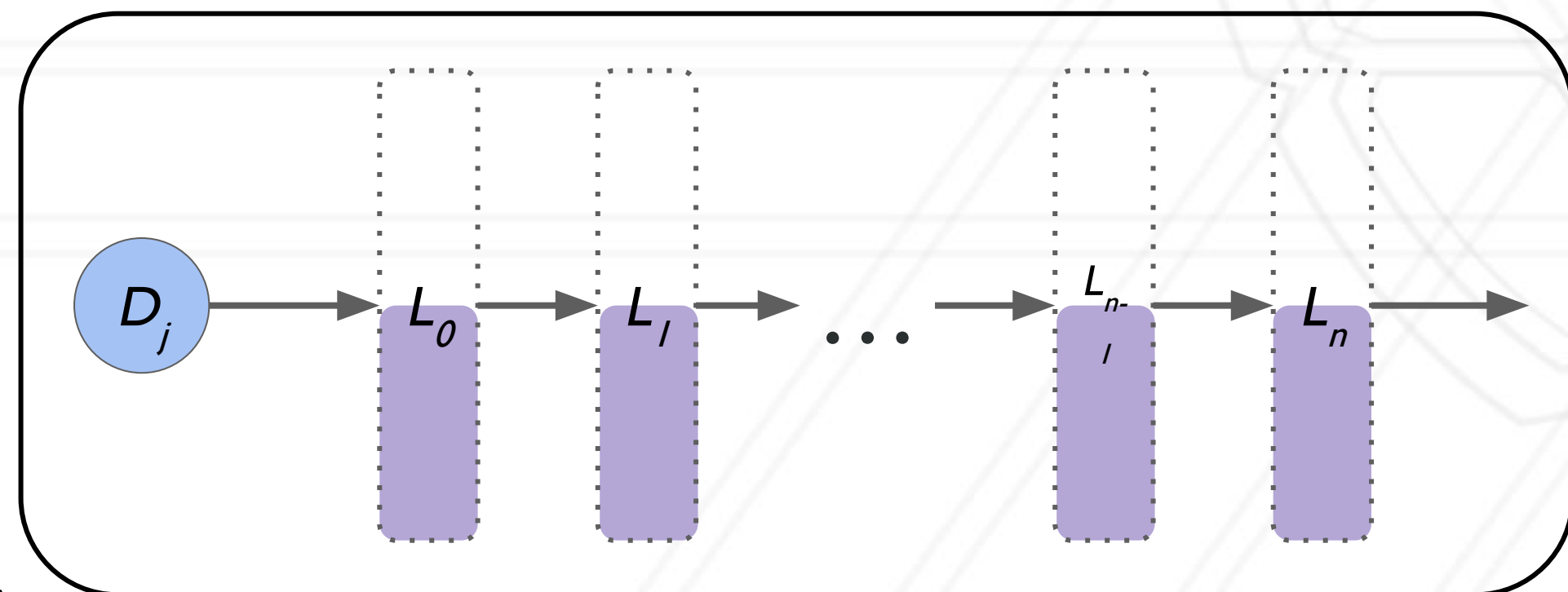
GPU 0



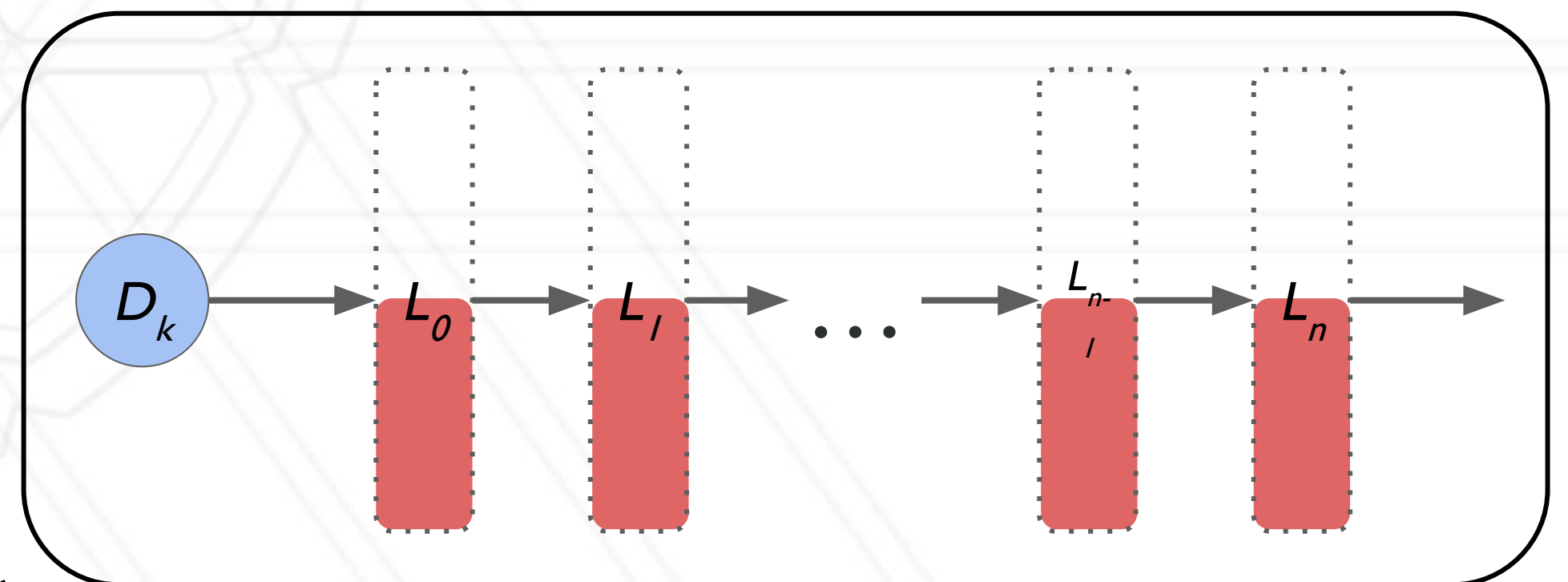
GPU 2



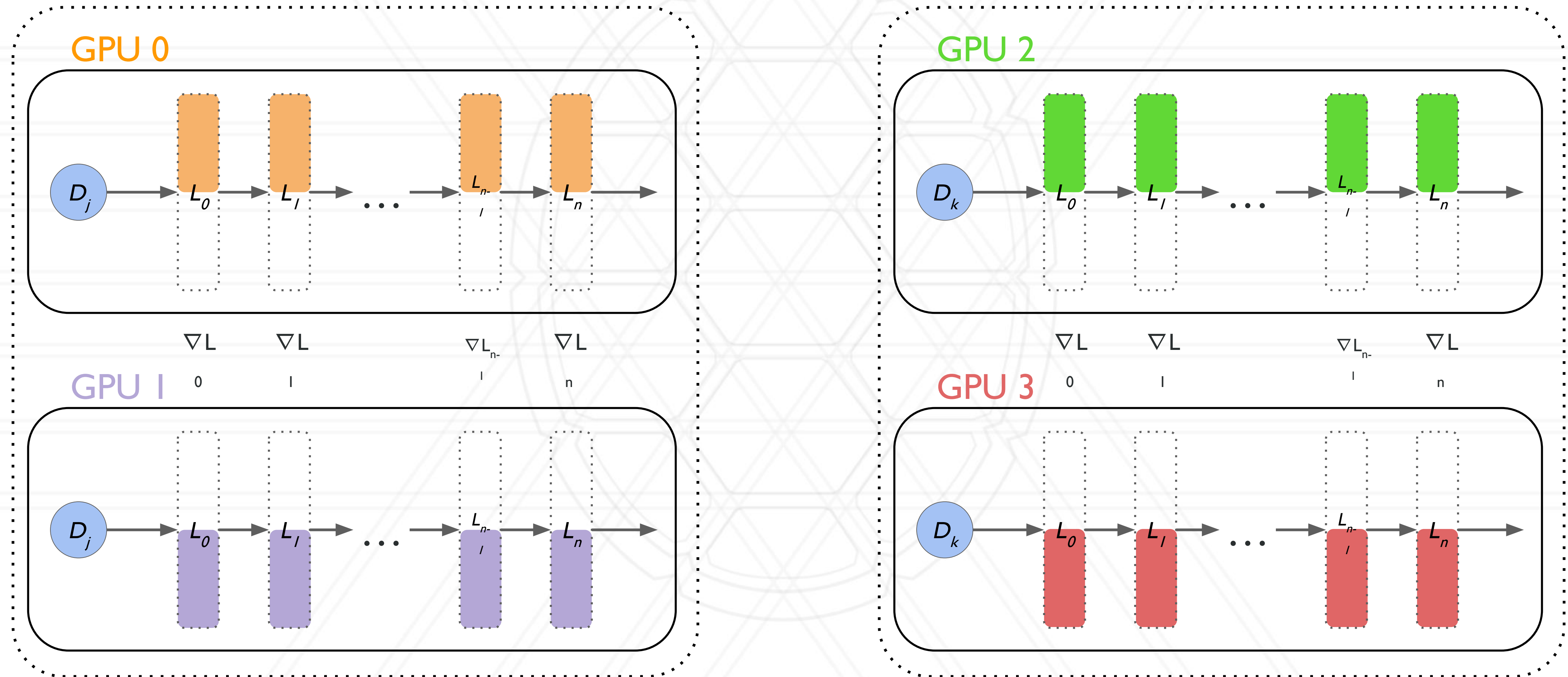
GPU 1



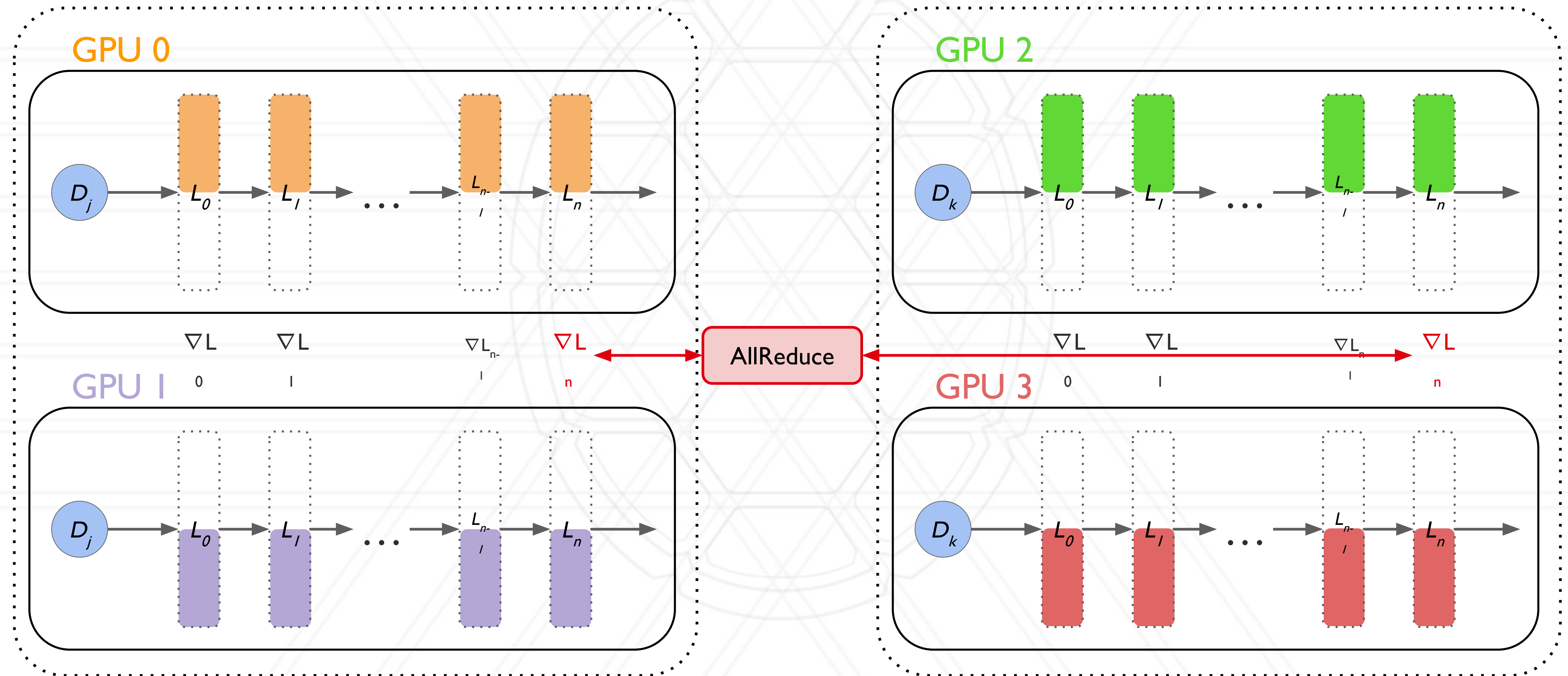
GPU 3



# Tensor + Data Parallel Training

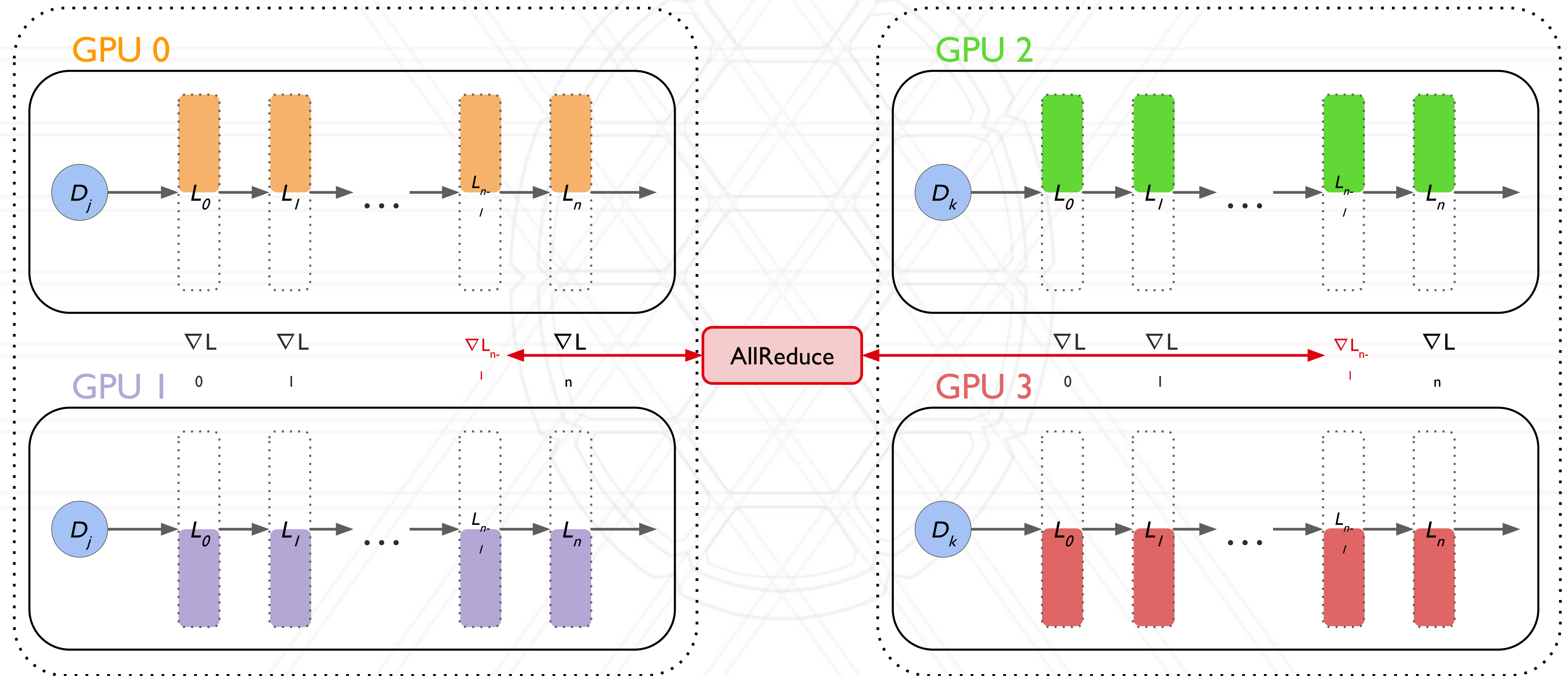


# Tensor + Data Parallel Training



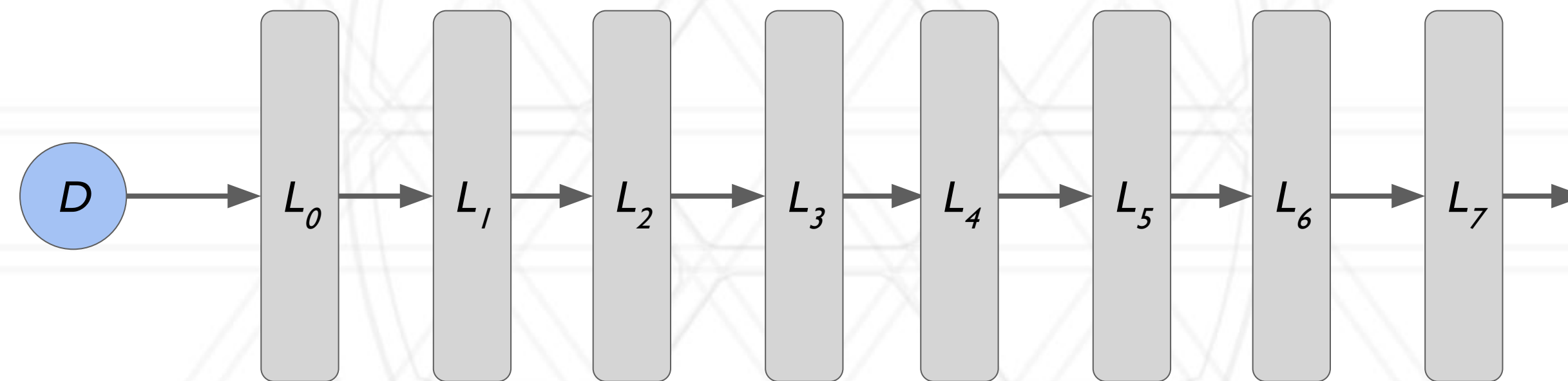


# Tensor + Data Parallel Training



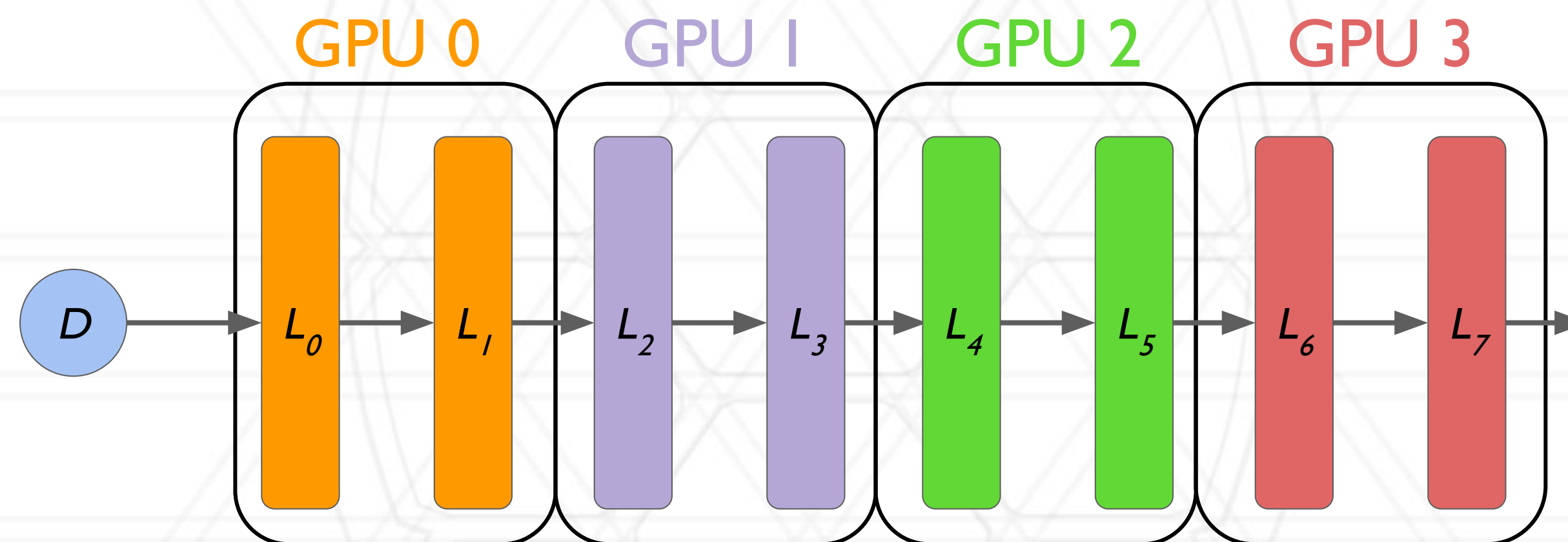
# Pipeline Parallel Training

---



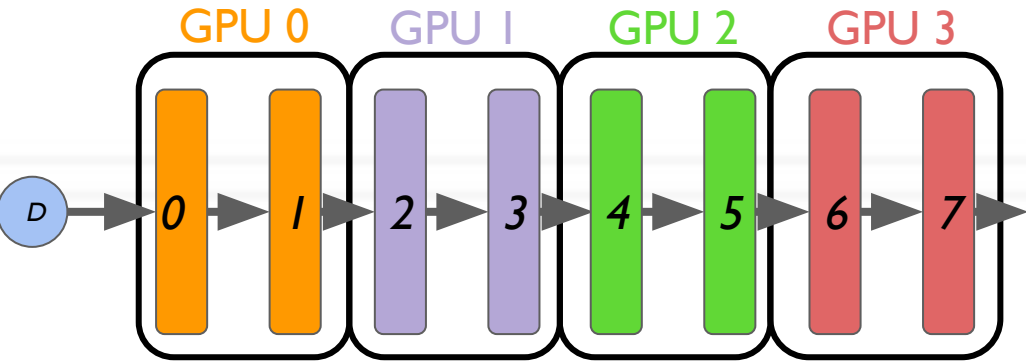
# Pipeline Parallel Training

---

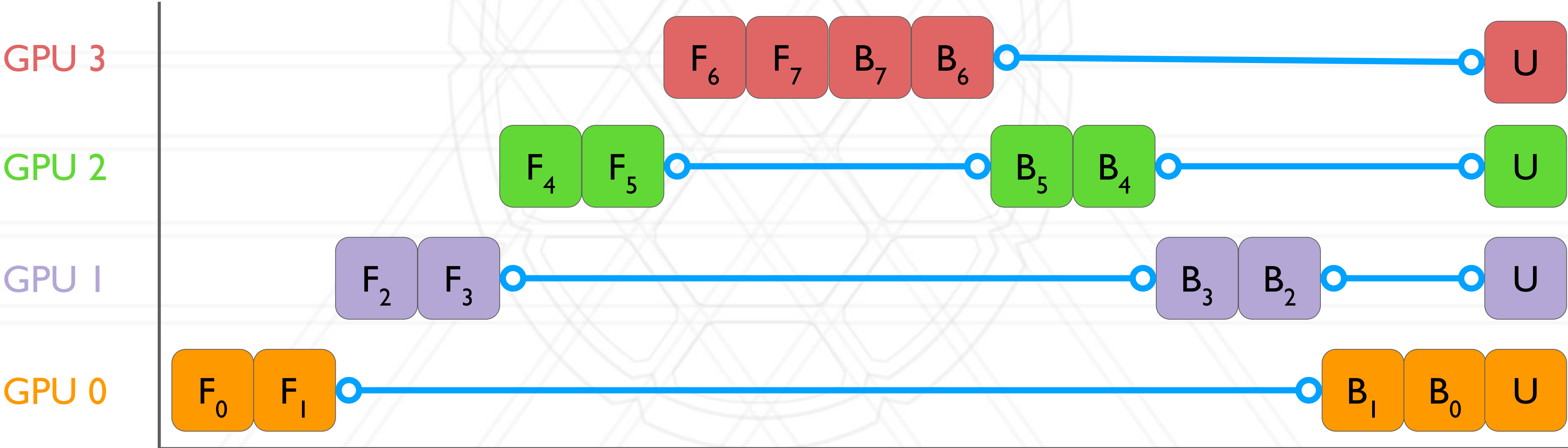




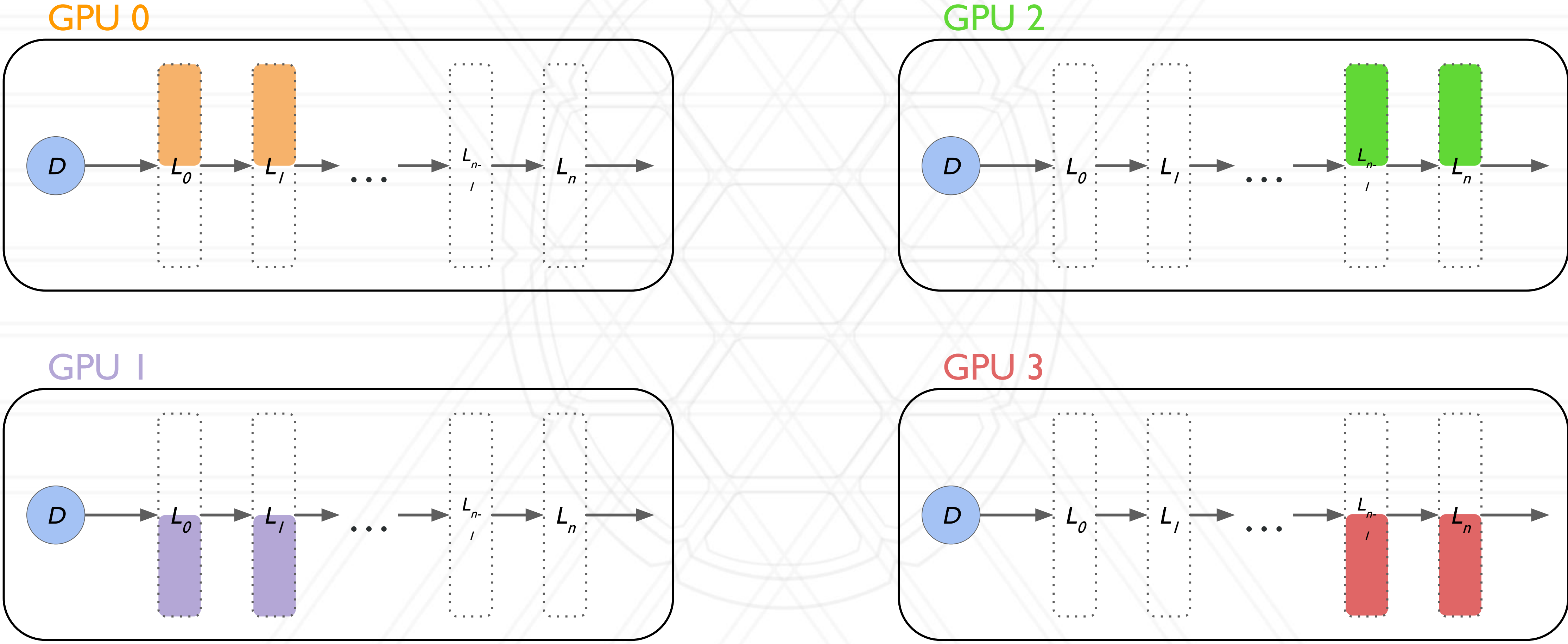
# Pipeline Parallel Training



There's a lot of research into filling in these "bubbles"



# Pipeline + Tensor Parallel Training



# Pipeline + Tensor + Data Parallel Training







UNIVERSITY OF  
MARYLAND