Q3. Part 1

1) minibatches per epoch = $\frac{76800}{1536} = 50$

total mini batches = $50 \times 10 = 500$

how many times is it distributed among 6 machines = $\frac{500}{6} = 83.\dot{3}$

Typical training round with an even load
= 1.5s (comms from master) + 4s (computation) + 1.5s (comms to master)
+ $6 \times 0.015_s = 7.09_s$

Training with the last 1/3 load = " " + $2 \times 0.015_s$
= 7.03

Total time = $7.09 \times 83 + 7.03 = 595.5_s$

2) Comm and agg time = $1.5_s + 1.5_s + 0.015m$
Training time = 4_s

$3 + 0.015m > 4$

$0.015m > 1$

$m > \frac{1}{0.015}$

$m > 66.\dot{6}$

$m \geq 67$

Q3. Part 2

Data per device = $\frac{204800}{8} = 25600$

Processed data per device after 10 epochs = 256000

batches processed = $\frac{256000}{32} = 800$

Assuming the updates are sent back after all 10 epochs are processed.

Total time = 7s (receiving the model) + 4 × 800 (updates) + 7s (sending the model back) = 3201 4s

Q2. Part 1

Assuming "the same" means equal number of operations

Asynchronous

Core 1: starts at $t = 0.15s$ Update intervals = 0.15 - 0.35

Core 2: at $t = 0.20s$ = 0.2 - 0.4

Core 3: at $t = 0.25s$ = 0.25 - 0.45

Core 4: at $t = 0.50s$ = doesn't update

Synchronous

Core 1: 0.0 - 0.2s 0.2s - 0.4s 0.4s - 0.6s with no delay in between

---

$$x_{async} = x_0 - \eta \nabla f_1(x_0) - \eta \nabla f_2(x_0) - \eta \nabla f_3(x_0)$$

$$x_{seq} = x_0 - \eta \nabla f_1(x_0) - \eta \nabla f_2(x_1) - \eta \nabla f_3(x_2)$$

where $x_1 = x_0 - \eta \nabla f_1(x_0)$ and $x_2 = x_1 - \eta \nabla f_2(x_1)$

$$\|x_{async} - x_{seq}\| = \| -\eta \nabla f_2(x_0) - \eta \nabla f_3(x_0) - (-\eta \nabla f_2(x_1) - \eta \nabla f_3(x_2)) \|$$

$$= \| -\eta \nabla f_2(x_0) - \eta \nabla f_3(x_0) + \eta \nabla f_2(x_1) + \eta \nabla f_3(x_2) \|$$

$$= \| (\eta \nabla f_2(x_1) - \eta \nabla f_2(x_0)) + (\eta \nabla f_3(x_2) - \eta \nabla f_3(x_0)) \|$$

triangle inequality $\leq \| \eta \nabla f_2(x_1) - \eta \nabla f_2(x_0) \| + \| \eta \nabla f_3(x_2) - \eta \nabla f_3(x_0) \|$

triangle inequality $\leq \| \eta \nabla f_2(x_1) \| + \| \eta \nabla f_2(x_0) \| + \| \eta \nabla f_3(x_2) \| + \| \eta \nabla f_3(x_0) \|$

(via $0 < \|\nabla f_i\| \leq \Delta$ $\leq \eta\Delta + \eta\Delta + \eta\Delta + \eta\Delta$

$\leq 4\eta\Delta$

Assuming "the same" means executed for the same amount of time 0.6s

Asynchronous

Execution times: Core 1: 0.15 - 0.35 0.35 - 0.55
Core 2: 0.2 - 0.4 0.4 - 0.6
Core 3: 0.25 - 0.45
Core 4: Doesn't update

Synchronous

Same as before.

[redacted]

Unsure of the notation but basically two more operations that don't cancel out with other operations so $4\eta\Delta + 2\eta\Delta = 6\eta\Delta$

Q2. Part 2

All gradient calcs finish at 0.5s

$t = 0.5_s + 0.3_s$ (aggregation) + 0.25_s (update) [redacted]
$= 1.05_s$