



大象、老鼠、森林：基於機器學習的雙時間尺度負載平衡器

A Two-Timescale Load Balancer Guided by ML

專題生：石浚邑、石維廉

指導教授：張燕光



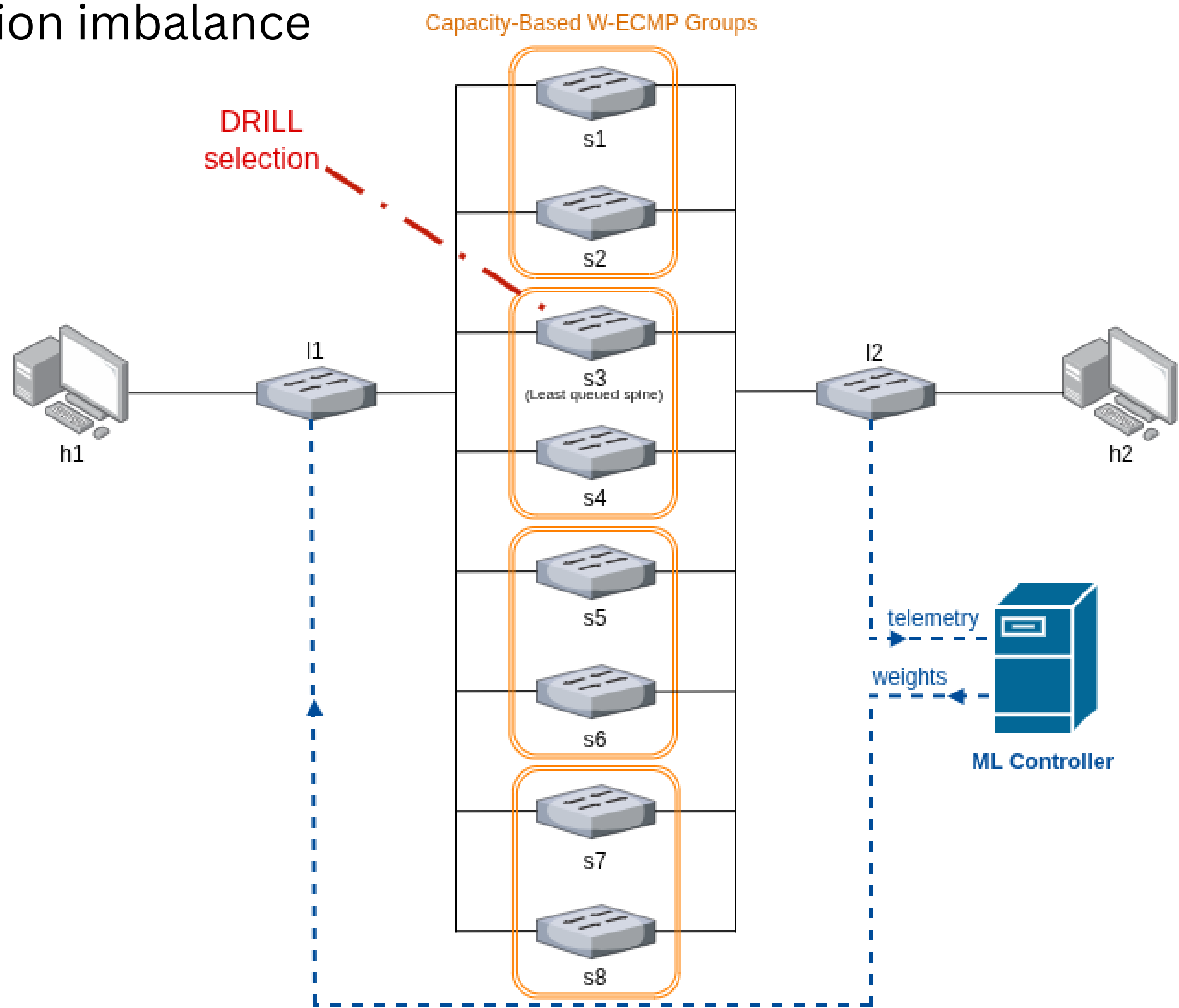
研究動機與目的：

當前的低延遲資料中心網路中，elephant / mice traffic 會造成短時間突發壅塞，尤其在非對稱 topology 中，傳統 ECMP / W-ECMP 容易產生負載不均，且 hash 分流後無法搬移 elephant flow 並根據 queue 狀態修正路徑。DRILL 能快速避開壅塞，卻缺乏全域平衡能力。

本專題提出雙時間尺度負載平衡方法：

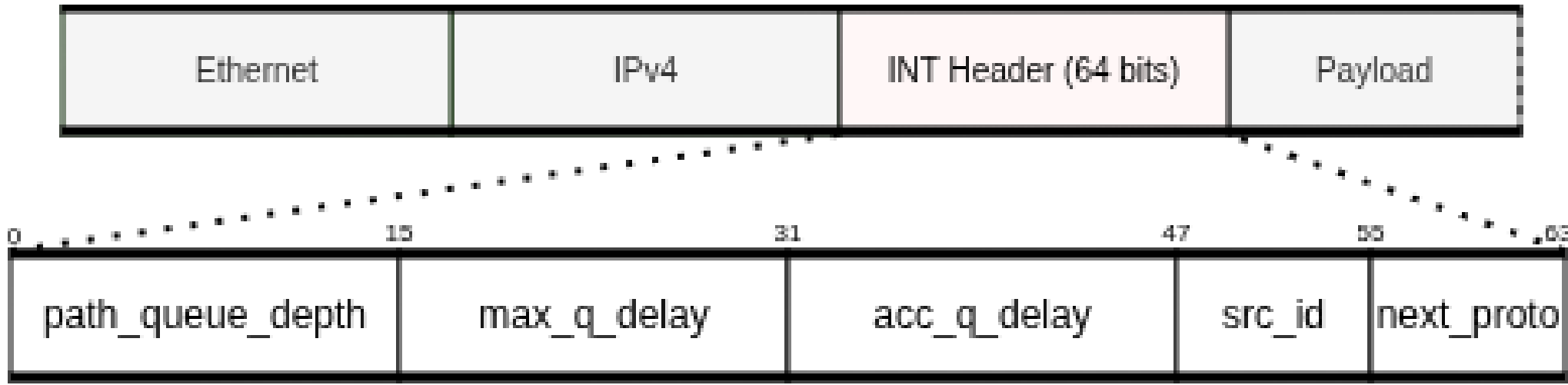
- 快時間尺度：DRILL 根據即時 queue 狀態進行封包層級選路
- 慢時間尺度：ML controller 每秒根據 telemetry 預測壅塞，調整 W-ECMP 權重

目標：降低 p95 / p99 tail latency、提升 throughput、降低 spine utilization imbalance



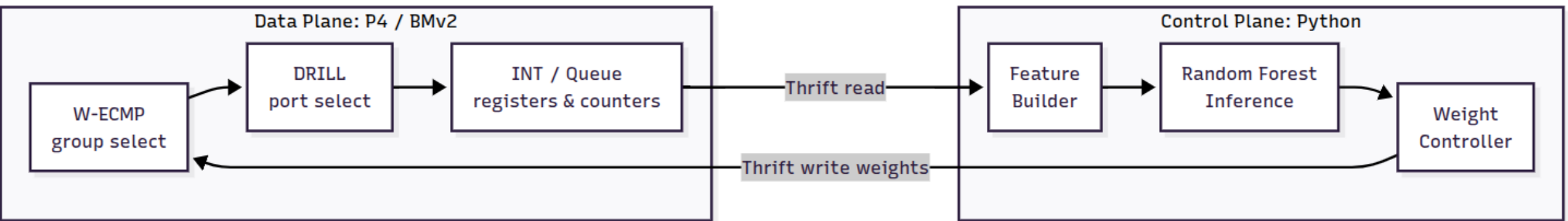
實作技術與環境：

- 使用工具與環境：Mininet、P4-Utills、BMv2 軟體交換器、scikit-learn Random Forest、P4、Python
- 資料收集：
 - P4 解析器支援自訂遙測標頭，EthernetType = 0x9999
 - 封包進入監測路徑時插入 telemetry header
 - 沿途累積 path queue depth、max queue delay、accumulated queue delay
 - 最後一跳將 telemetry 結果寫入 P4 registers
 - Python controller 每秒讀取 registers，建立模型輸入特徵



- 機器學習模型與控制流程：
 - 將 raw port-level telemetry 轉換為 topology-independent features
 - Random Forest 預測 next-second latency / loss / anomaly
 - Controller 根據預測結果調整 W-ECMP weights
 - 實驗比較 ECMP、W-ECMP、Naive DRILL 與 ML Hybrid

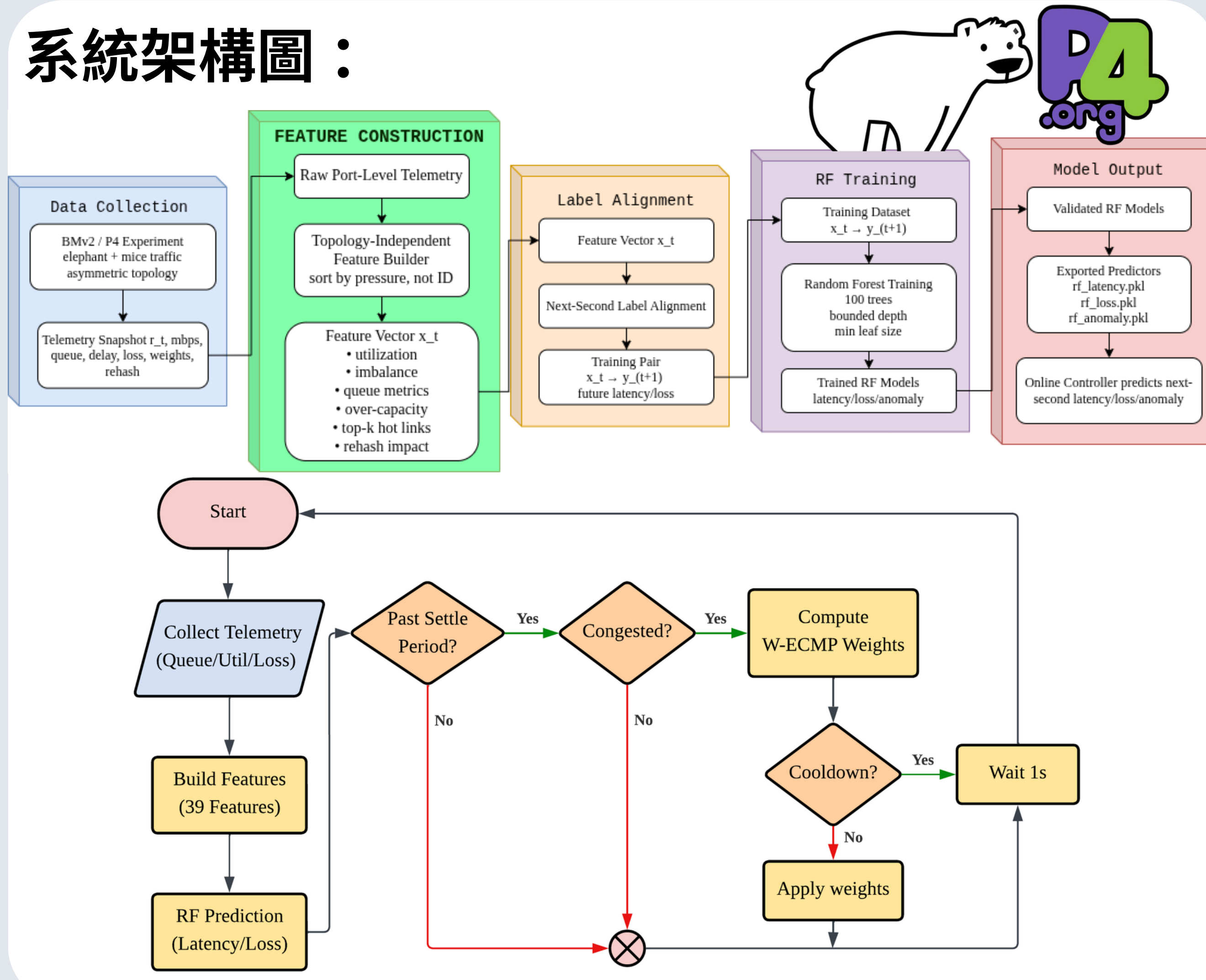
方法	控制尺度	主要功能	限制
W-ECMP	流層級	依容量比例分流	路徑選定後不會中途修正
Naive DRILL	封包層級	依即時佇列選路	不感知鏈路容量差異
ML Hybrid	雙尺度	DRILL 與 W-ECMP + ML 權重修正	需要遙測資料與模型推論



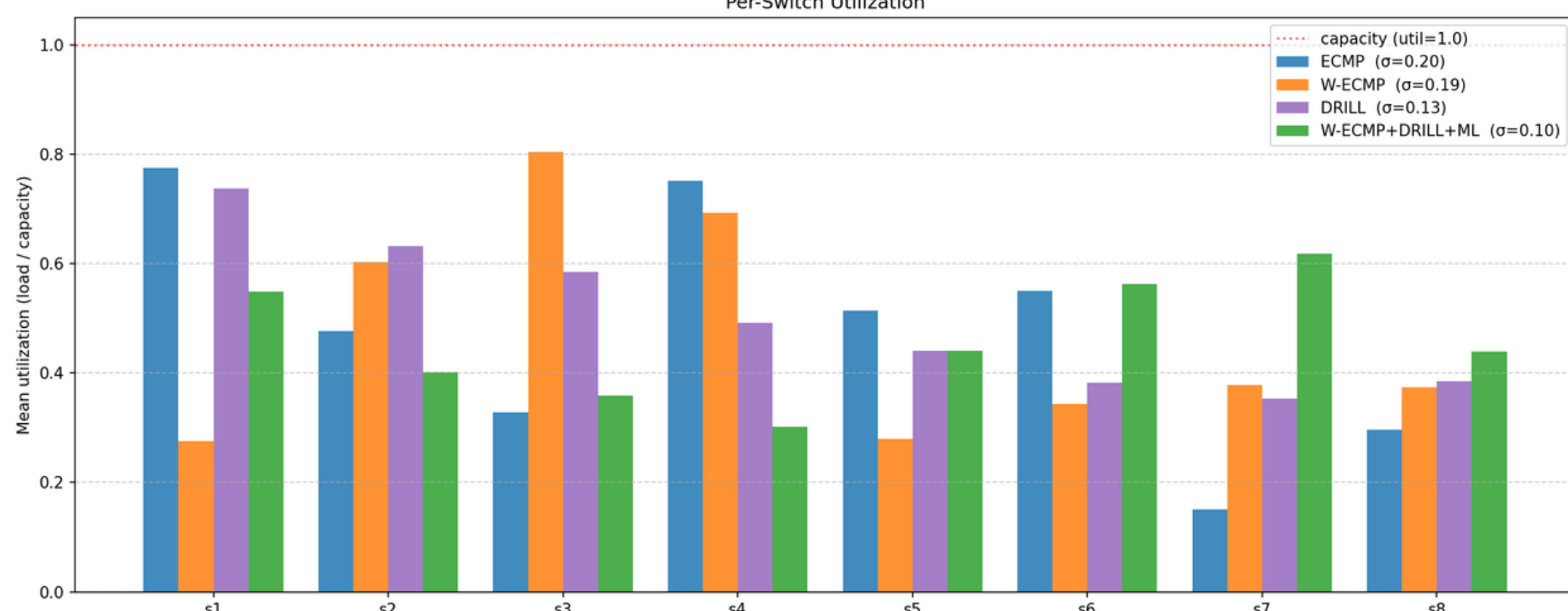
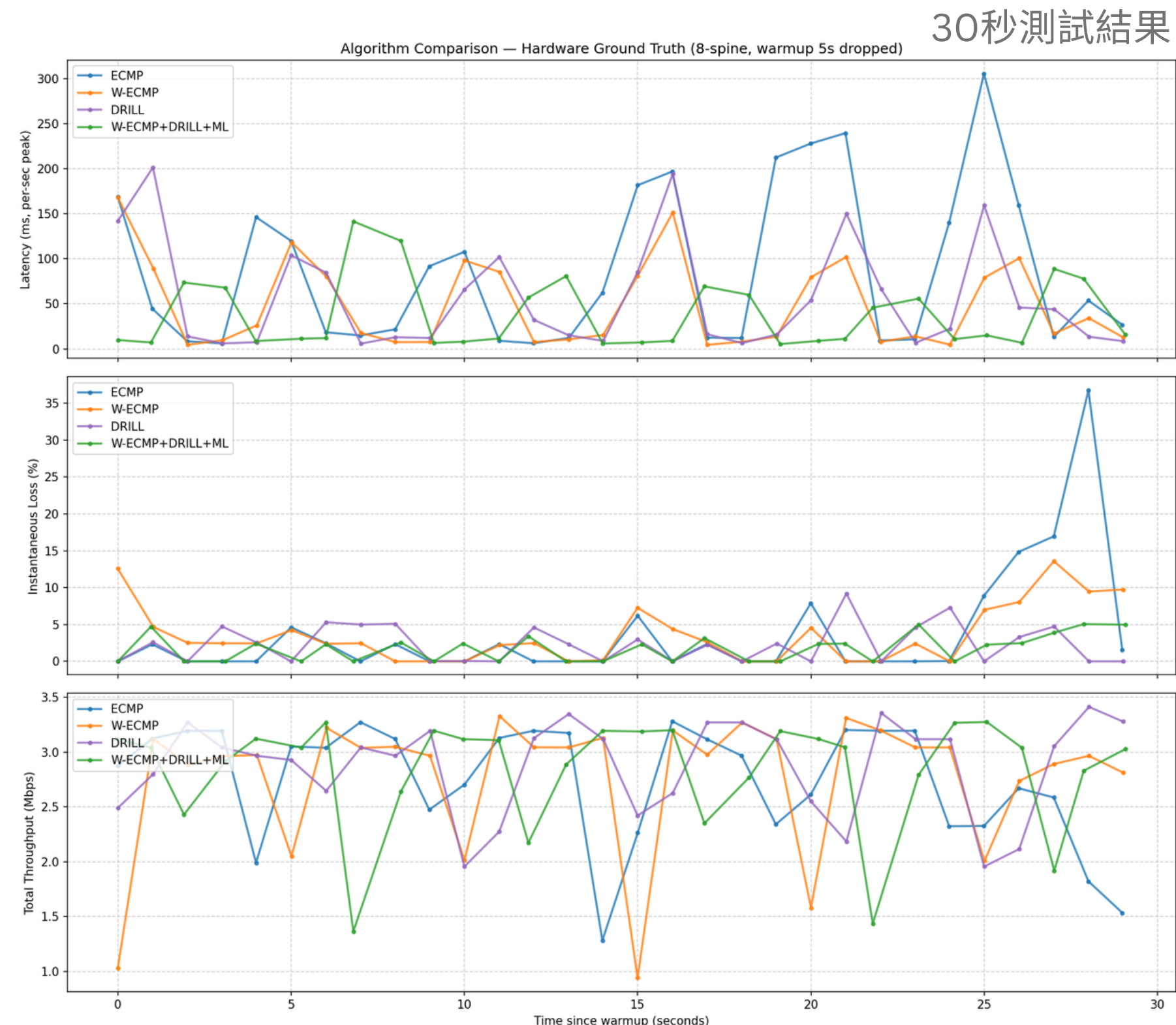
ML Hybrid 雙尺度下的負載平衡：

快尺度的 DRILL 反應快，在群組內封包導向負載低的 switch。慢尺度的 ML+W-ECMP 預測群組層級的持續失衡，透過調整權重把整條 elephant flow 的流量搬走。一快一慢搭起來，演算法從封包到流的失衡都涵蓋住。

系統架構圖：



成果展示：



演算法	Lat p50	Lat p95	Lat p99	Throughput	Util σ
ECMP	108.66 ms	443.84 ms	1185.99 ms	2.53 mbps	0.1180
W-ECMP	79.64 ms	1025.95 ms	2485.76 ms	2.41 mbps	0.0870
Naive DRILL	18.97 ms	159.50 ms	187.78 ms	2.87 mbps	0.1220
ML Hybrid	20.27 ms	141.16 ms	172.24 ms	2.86 mbps	0.0430

此演算法降成功降低 p99 與 p95 延遲並達到最有效的負載平衡。

結論與未來展望：

實驗結果顯示，ML Hybrid 在非對稱 topology 與 elephant / mice traffic 下，能有效降低 p95 / p99 tail latency 並改善 spine utilization imbalance。

不過最重要的是「機器學習應用於資料中心負載平衡」的可能性。若未來面對更真實的資料中心的複雜流量，則可在模型中加入更多高階特徵，例如：時間週期、連線數、flow duration、任務類型與應用層負載資訊等。對於具週期性的負載，模型可學習其變化模式並提前預測壅塞，利用更多語意特徵來提前辨識即將形成的熱點並預先平衡。機器學習使資料中心網路控制從被動反應推向主動預防。