A PROOFS

A.1 Proof of Proposition 1

Proof. Let $\psi_i(v)$ represent the disk usage of node n_i at timestamp v. For $v < \mu \vee \mu + \text{TTL} < v$, since $\psi_i(v) = \text{TTL} \times C$, we have $\mathcal{D}(t_v) \equiv 0$. Define $z_i(t) \in \{0, 1\}$, where $z_i(t) = 1$ indicates that node n_i is part of the cluster at timestamp t, and $z_i(t) = 0$ otherwise. Thus, we have $\psi_i(v) = \int_{v-TTL}^{v} C \times z_i(t)\, dt$. Let $\overline{\psi}(v)$ denote the average disk usage of all nodes at timestamp v. It follows $\overline{\psi}(v) = \frac{C}{\alpha+\beta} \times (\alpha \times \text{TTL} + \beta \times (v - \mu))$. Let $\mathcal{N}$ represent the set of all cluster nodes after expansion. The std of disk usage at timestamp v is given by $\mathcal{D}(v) = \sqrt{\frac{1}{\alpha+\beta} \times \sum_{n_i \in \mathcal{N}} \left(\psi_i(v) - \overline{\psi}(v)\right)^2}$.

Let $\mathcal{N}_\alpha$ denote the set of α nodes that existed before the cluster expansion, and let $\mathcal{N}_\beta$ represent the set of β newly expanded nodes. Then, for all $n_i \in \mathcal{N}_\alpha$, $\psi_i(v) - \overline{\psi}(v) = \frac{\beta C}{\alpha+\beta} \times (\mu + \text{TTL} - v)$, and for all $n_j \in \mathcal{N}_\beta$, $\psi_j(v) - \overline{\psi}(v) = -\frac{\alpha C}{\alpha+\beta} \times (\mu + \text{TTL} - v)$. Consequently, Equation 2 is derived by combining all the equations above. □

A.2 Proof of Theorem 1

Proof. Given an instance, the scatter width increment $\Sigma' - \Sigma$ can be easily computed by adding edges and comparing the vertices' degree before and after this operation. Next, we construct a reduction from the independent set decision problem [18], defined as $ISDP(\mathcal{V}, \mathcal{E}, \rho)$, which asks whether there exists an independent set of size at least ρ in the given undirected graph $G = (\mathcal{V}, \mathcal{E})$.

Let $\dot{\mathcal{N}} = \mathcal{V}, \dot{\rho} = \rho$, the adjacency matrix $\dot{\mathcal{A}}$ can be converted from $\mathcal{E}$ in polynomial time. We construct $\dot{\mathcal{W}} = [0]_{1\times n}, \omega = n$ and $\dot{m} = 1$, indicating that only 1 scheme needs to be found while all vertices are available for inclusion. Let $\dot{\Lambda} = \rho(\rho - 1)$, we have $ISDP \leq_p RPDP$ because there exists an independent set of size at least ρ in $G = (\mathcal{V}, \mathcal{E})$ iff. there exists an "optimal" scheme that increases $\rho - 1$ scatter width for each included node. □

A.3 Proof of Lemma 1

Proof. In Algorithm 1, the $\lfloor \frac{\rho}{2} \rfloor$ nodes that hold the fewest replicas are included when selecting $\mathbf{n}_\alpha$ to ensure storage balance. Thus, generating $\frac{n}{\lfloor \frac{\rho}{2} \rfloor}$ good placement schemes ensures that each node is included in a good placement scheme at least once. □

A.4 Proof of Lemma 2

Proof. Let $g_{\alpha^*} = \sum_{n_\gamma \in \mathbf{n}^*_\alpha} g\left(n_\gamma, \mathcal{N}_i \setminus \mathbf{n}^*_\alpha, \mathcal{A}\right)$, which is equal to the sum of complement degrees from $\mathbf{n}^*_\alpha$ to the remainder of nodes in $\mathcal{N}_i$. The initial complement degree for any $n_\gamma \in \mathbf{n}^*_\alpha$ under $\mathcal{A}$ is $|\mathcal{N}_i| - \lfloor \frac{\rho}{2} \rfloor$. Additionally, each time a vertex is included in $\mathbf{n}_\alpha$, its complement degree under $\mathcal{A}$ decreases by at most $\lfloor \frac{\rho}{2} \rfloor$, and it can be selected into $\mathbf{n}_\alpha$ at most ω times. Thus the lower bound is $g_{\alpha^*} \geq \lfloor \frac{\rho}{2} \rfloor \left(|\mathcal{N}_i| - \lfloor \frac{\rho}{2} \rfloor (1 + \omega)\right)$. According to the pigeonhole principle, when $g_{\alpha^*} > \left(\lfloor \frac{\rho}{2} \rfloor - 1\right)\left(|\mathcal{N}_i| - \lfloor \frac{\rho}{2} \rfloor\right)$, there must exist a $n_\iota \in \mathcal{N}_i$ s.t. $g(n_\iota, \mathbf{n}^*_\alpha, \mathcal{A}) = \lfloor \frac{\rho}{2} \rfloor$. Finally, this Lemma is proven by combining the two inequations above. □

A.5 Proof of Lemma 3

Proof. Let $g_\alpha = \sum_{n_\gamma \in \mathbf{n}_\alpha} g\left(n_\gamma, \mathcal{N}_j, \mathcal{A}\right)$, which is equal to the sum of complement degrees from $\mathbf{n}_\alpha$ to all nodes in $\mathcal{N}_j$. Following the same processes as Lemma 2, we obtain an inequation for the lower bound $g_\alpha \geq \left(\lfloor \frac{\rho}{2} \rfloor + 1\right)\left(|\mathcal{N}_j| - \left(\lfloor \frac{\rho}{2} \rfloor + 1\right)\omega\right)$ and another inequation for the pigeonhole principle $g_\alpha > \lfloor \frac{\rho}{2} \rfloor |\mathcal{N}_j|$. This lemma is proven by combining the two inequations above. □

A.6 Proof of Lemma 4

Proof. In the worst case, generating ω schemes fills ρ nodes, as though all ω schemes select the same nodes. Since Algorithm 1 always avoids selecting duplicated node combinations to maximize the scatter width, the rate at which the number of filled nodes increases will not be faster than the aforementioned worst case. □

A.7 Proof of Propositions 2

Proof. Combining Lemmas 2 and 4, the lower bound for the number of good schemes that Algorithm 1 can generate is $\check{r} \geq (n - \omega + 1) \times \frac{\omega}{2}$. Taking this inequation and applying it to Lemma 1, we obtain $\check{\sigma}_n \geq \frac{n-\omega+1}{2n}$, where the limit is $\frac{1}{2}$. □

A.8 Proof of Proposition 3

Proof. Integrating Lemmas 2 and 3 derives the sufficient condition that for a good scheme to exist—$\forall 0 \leq i < \lceil \frac{\rho}{2} \rceil, |V_i| > \left(\lfloor \frac{\rho}{2} \rfloor + 1\right)^2 \omega$. Since $\left(\lfloor \frac{\rho}{2} \rfloor + 1\right)^2 \omega > \lfloor \frac{\rho}{2} \rfloor^2 \omega - \lfloor \frac{\rho}{2} \rfloor$ works $\forall \omega > 0$, this condition applies universally. Then following the same steps as Proposition 2 and incorporating the inequations that $\frac{\rho-1}{2} \leq \lfloor \frac{\rho}{2} \rfloor \leq \frac{\rho}{2} \leq \lceil \frac{\rho}{2} \rceil$, we obtain $\check{\sigma}_n \geq \frac{4\min(|\mathcal{N}_i|) - (\rho+1)^2}{16n}$. As $|\mathcal{N}_i|$ approximates $\frac{n}{\lceil \frac{\rho}{2} \rceil}$, the limit for $\check{\sigma}_n$'s lower bound is $\frac{1}{4 \times \lceil \frac{\rho}{2} \rceil}$. □

A.9 Proof of Proposition 4

Proof. As long as a group of nodes containing any employed replica placement scheme fail simultaneously, at least one shard becomes unavailable. Assuming that the cluster has employed r pairwise distinct placement schemes. Given that the number of schemes for selecting ρ nodes from n nodes to place a shard is $\binom{n}{\rho}$, if ρ nodes fail concurrently, the probability that the combination of these ρ nodes encounters an employed scheme is $\epsilon = \frac{r}{\binom{n}{\rho}}$.

In the case where m nodes fail simultaneously, these failed nodes correspond to $\binom{m}{\rho}$ possible schemes. Since each has ϵ to be employed, the probability $P(X(n, m) = k)$ follows binomial distribution $Binomial(\binom{m}{\rho}, \epsilon)$, i.e., $P(X(n, m) = k) = \binom{\binom{m}{\rho}}{k} \epsilon^k (1 - \epsilon)^{\binom{m}{\rho} - k}$. Subsequently, given that the n is relatively large and ϵ is small in real-world deployments, and considering the applicability of the Poisson distribution to such rare events, we apply the Poisson limit theorem [13] to approximate the binomial distribution. The rate parameter for $Poisson(\lambda)$ is $\lambda = \binom{m}{\rho} \times \epsilon$. Therefore, as the number of nodes deployed in the cluster gradually expands, we obtain $\lim_{n\to\infty} P(X(n, m) = k) = e^{-\lambda} \times \frac{\lambda^k}{k!}$. Finally, this proposition is proved by the fact that $P(X(n, m) \geq 1) = 1 - P(X(n, m) = 0)$. □

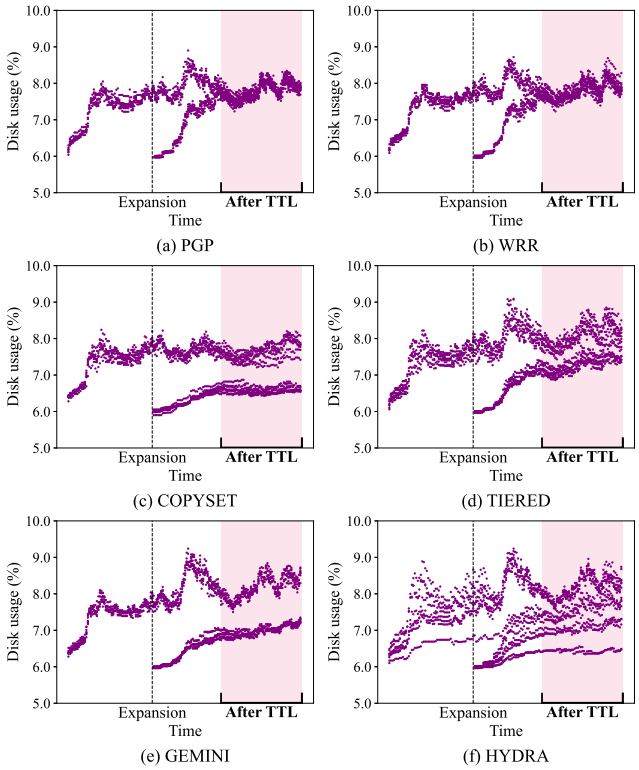

Figure 19: Disk usage distribution in the expansion scenario.

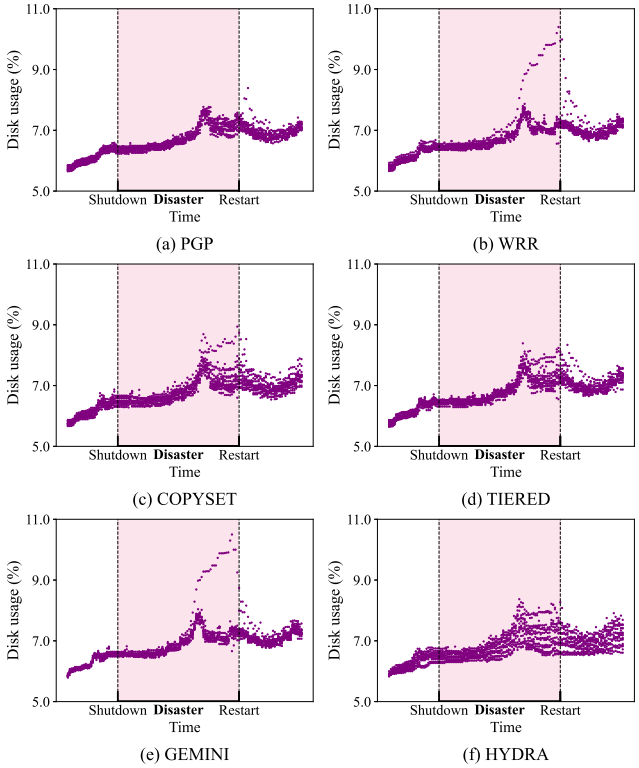

Figure 20: Disk usage distribution in the disaster scenario.

Table 4: Average throughput standard deviation generated through different leader selection algorithms.

Scenario	Algorithm	STD (K points/s)	Difference
After expansion (Figure 15(b))	**CFS** (ours)	**10.541**	-
	GREEDY	44.573	+322.9%
	ESDB	78.999	+649.4%
	LOGSTORE	138.347	+1212.5%
	RANDOM	106.827	+913.4%
During recovery (Figure 18(b))	**CFS** (ours)	**13.528**	-
	GREEDY	21.683	+60.3%
	ESDB	24.624	+82.0%
	LOGSTORE	31.710	+134.4%
	RANDOM	39.511	+192.1%

Table 3: Maximum disk usage standard deviation generated through different replica placement algorithms.

Scenario	Algorithm	STD (%)	Difference
Expansion after TTL (Figure 10)	**PGP** (ours)	**0.153**	-
	WRR	0.172	+12.4%
	COPYSET	0.680	+344.4%
	TIERED	0.533	+248.4%
	GEMINI	0.820	+435.9%
	HYDRA	0.837	+447.1%
During disaster (Figure 13)	**PGP** (ours)	**0.233**	-
	WRR	0.803	+244.6%
	COPYSET	0.538	+130.9%
	TIERED	0.396	+70.0%
	GEMINI	0.804	+245.1%
	HYDRA	0.505	+116.7%

A.10 Proof of Lemma 5

Proof. According to Hall's marriage theorem [20], if for any subset of the shard set $\mathcal{R}$, the sum of the number of leader replicas that can be held by its associated nodes exceeds its size, each shard can be matched with a node. Since the commonly configured replication factor $\rho > 1$, which means for any shard, the sum of the number of leader replicas that can be held by its associated nodes is $\rho > 1$, the above condition is satisfied. □

A.11 Proof of Proposition 5

Proof. Lemma 5 guarantees that $\forall r_k \in R, p_k \in r_k$ holds in every generated leader distribution. The storage balance constraint of Algorithm 1 leads to a balanced replica distribution $\forall n_i \in \mathcal{N}, \omega - 1 \leq \omega_i \leq \omega$. Hence the remaining proof is to show that the final distribution is optimally balanced iff. $\forall n_i \in \mathcal{N}, \left\lfloor \frac{\omega}{\rho} \right\rfloor \leq \eta_i \leq \left\lceil \frac{\omega}{\rho} \right\rceil$.

Necessarily, $\forall n_i \in \mathcal{N}, \left\lfloor \frac{\omega}{\rho} \right\rfloor \leq \eta_i \leq \left\lceil \frac{\omega}{\rho} \right\rceil$ results in the optimally balanced leader distribution. Since the cost for a node n_i to own η_i

leaders—$\sum_{k=1}^{\eta_i} \delta(k) = \eta_i^2$—is a convex function, and it is additive, combining this property with Jensen's inequality [23], when $\forall n_i \in \mathcal{N}$, $\left\lfloor \frac{\omega}{\rho} \right\rfloor \leq \eta_i \leq \left\lceil \frac{\omega}{\rho} \right\rceil$, the total cost of $\widetilde{G}$ is minimized.

Sufficiency obtained from the construction process of the next augmenting path for selecting the leader replica of shard r_i. To present vertices involved in this path, we define the relation $\sim$ such that two vertices $v_i \sim v_j$ if, when considering only edges $e \in \widetilde{B}$, v_i and v_j are connected. Let $\hat{\mathcal{R}} = \{r_j : r_i \sim r_j\}$ and $\hat{\mathcal{N}} = \{n_j : r_i \sim n_j\}$. The next augmenting path must consist of $S \rightarrow r_i \rightarrow \cdots \rightarrow n_j \rightarrow T$, where the number of leader replicas η_j possessed by vertex n_j is minimum among $\forall n_\iota \in \hat{\mathcal{N}}$, ensuring that the cost of this path, $2\eta_j + 1$, is minimized. As each augmenting path ends with reaching the sink vertex T through the node that owns the least number of leader replicas, $\forall n_i \in \mathcal{N}$, $\left\lfloor \frac{\omega}{\rho} \right\rfloor \leq \eta_i \leq \left\lceil \frac{\omega}{\rho} \right\rceil$ holds eventually. □