UDC 004.75

FORMAL MODEL OF EDGE–FOG–CLOUD-BASED DATA DISTRIBUTION

A. Klychev

Oguz han Engineering and Technology University of Turkmenistan, Ashgabat, Turkmenistan,
annamyrat.gylyjov@etut.edu.tm

Abstract: smart-city and IoT systems generate heterogeneous data whose latency, sensitivity and computational demands cannot be met efficiently by a single computing tier. This paper formalizes data distribution across a hierarchical Edge–Fog–Cloud architecture, represented as a labeled graph whose nodes are partitioned into edge, fog and cloud layers with distinct capacities and latency ranges. A layer-assignment function maps each data unit to a processing layer according to its latency requirement, sensitivity level, computational load and current resource utilization. Four categories of distribution criteria – temporal, semantic, security and resource – are formalized, together with a weighted multi-criteria score function and a dynamic adaptation mechanism, providing a theoretical basis for latency-aware, privacy-preserving data placement.

Keywords: edge computing; fog computing; cloud computing; formal model; data distribution; latency; IoT; distributed systems.

I. INTRODUCTION

The proliferation of Internet-of-Things (IoT) sensors, video cameras and actuators in smart-city infrastructures has produced data volumes and velocities that a purely centralized, cloud-based processing model can no longer accommodate. Long-distance transmission between IoT sources and remote data centers introduces propagation delay together with substantial bandwidth demand, which is incompatible with applications that require millisecond-scale reaction, such as traffic-signal control or fire-hazard detection. These limitations motivate a hierarchical Edge–Fog–Cloud architecture, in which processing is distributed across three tiers positioned at increasing distance from the data source and possessing increasingly large, but more remote, computational capacity.

While the architectural roles of the edge, fog and cloud tiers are well documented in the literature, a compact formal representation that simultaneously captures topology, per-layer capacity, latency bounds, data sensitivity and dynamic load is less common, and existing descriptions often remain qualitative. This paper addresses that gap by (i) representing the Edge–Fog–Cloud system as a labeled graph together with a data-assignment function, (ii) formalizing four categories of distribution criteria that determine the processing layer of a given data unit, and (iii) introducing a multi-criteria decision function and a state-dependent dynamic adaptation mechanism that together support latency-aware and privacy-aware data placement. The model is intended as a theoretical basis for the simulation and performance evaluation to be carried out in subsequent stages of this research.

II. FORMAL MODEL OF THE EDGE–FOG–CLOUD ARCHITECTURE

A. System and Data Representation

The Edge–Fog–Cloud system consists of three functionally distinct tiers. The edge tier comprises small, physically distributed computing units co-located with IoT devices; its principal role is first-stage filtering and local decision-making, so that typically only 5–15% of raw sensor data is forwarded to higher tiers. The fog tier aggregates data from a set of edge nodes belonging to a common geographic area, performs medium-scale analytics and coordinates resource allocation among the edge nodes it supervises; the concept was introduced by Cisco in 2012 [1] and later standardized by the IEEE as the OpenFog Reference Architecture (IEEE 1934-2018) [2]. The cloud tier provides elastic, large-scale computing and long-term storage for city-wide data integration, model training and cross-system analytics, at the cost of higher end-to-end latency and dependence on external infrastructure providers.

$$G = (N, E, L), \tag{1}$$

where N is the set of all computing nodes, $E \subseteq N \times N$ is the set of communication links between nodes, and $L = \{edge, fog, cloud\}$ is the set of layer labels. The node set is partitioned according to layer membership as $N = N_{edge} \cup N_{fog} \cup N_{cloud}$, with $N_{edge} \cap N_{fog} = N_{fog} \cap N_{cloud} = \emptyset$.

Data circulating in the system are represented as a set:

$$D = \{d_1, d_2, \ldots, d_n\}, \tag{2}$$

where each data unit d_i is characterized by a tuple ($|d_i|$, $type_i$, $L_{req}(d_i)$, $S(d_i)$, ts_i): $|d_i|$ is the data volume (kB), $type_i$ denotes the semantic data type, $L_{req}(d_i)$ is the maximum tolerable end-to-end latency (ms), $S(d_i) \in \{1, 2, 3, 4, 5\}$ is a discrete sensitivity level (1 = lowest, 5 = highest, e.g. personally identifiable or state-classified information), and ts_i is the generation timestamp. Four flow directions are distinguished within G: upward flow (device → edge → fog → cloud), downward flow (control and update messages traveling in the opposite direction), lateral flow (exchange between nodes of the same layer), and feedback flow (results returned to the originating source).

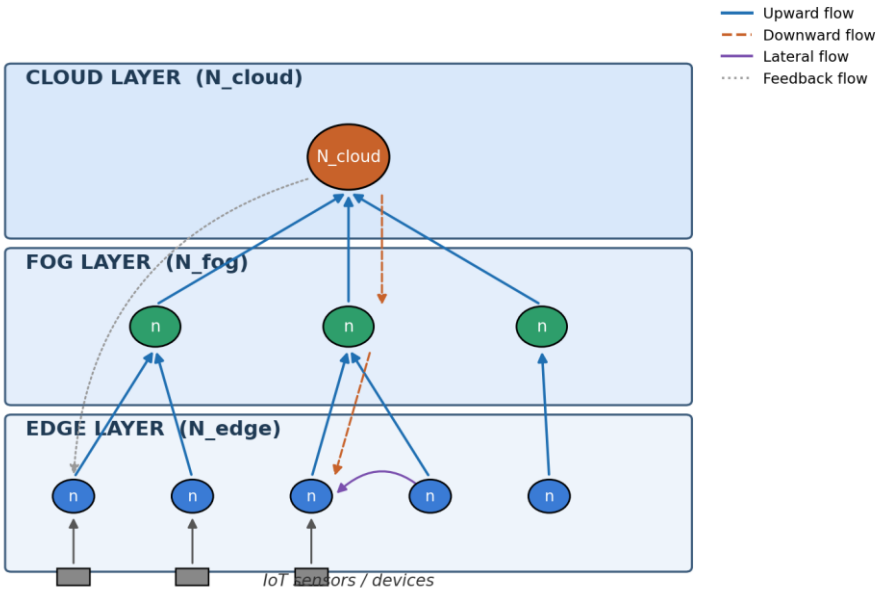


Figure 1. Layered Edge–Fog–Cloud architecture and the four data-flow directions (upward, downward, lateral, feedback)

Table 1. Comparative characteristics of the Edge, Fog and Cloud layers

Parameter	Edge	Fog	Cloud
Typical latency range	0.1–10 ms	10–100 ms	50–500 ms
Typical computing capacity	1–16 CPU cores, 1–32 GB RAM	8–64 CPU cores, 32–512 GB RAM	Elastic; up to thousands of CPU cores and GPUs
Typical storage capacity	Small local SSD/DRAM (short-term)	Medium-term local database / cache	Petabyte–exabyte scale, long-term
Primary function	Local filtering and real-time decision-making	Cross-node aggregation and coordination	Large-scale analytics, model training, long-term storage

B. Layer Capacity, Latency and Transfer-Time Constraints

Each layer $l \in L$ is characterized by a processing capacity C_l, expressed as available computational throughput, with the strict ordering

$$C_{edge} \ll C_{fog} \ll C_{cloud} \quad (3)$$

reflecting the trade-off between physical proximity and computational scale. Correspondingly, each layer operates within a characteristic latency range: $\tau_{edge} \in [0.1, 10]$ ms, $\tau_{fog} \in [10, 100]$ ms and $\tau_{cloud} \in [50, 500]$ ms. The choice of processing layer for a data unit d_i is captured by an assignment function

$$F: D \times N \rightarrow L, \quad (4)$$

which maps each data unit, given the current network state, to the layer responsible for its processing; F depends primarily on $L_{req}(d_i)$ and $S(d_i)$, and secondarily on available capacity and network load (Section II-C).

For a communication link $e(n_i, n\square) \in E$ connecting nodes n_i and $n\square$, let $B(e)$ denote its bandwidth and $\tau(e)$ its propagation latency. The time required to transfer a data unit d_i across e is:

$$T_{transfer}(d_i, e) = |d_i| / B(e) + \tau(e). \quad (5)$$

The overall latency experienced by d_i is the sum of propagation, processing and queuing delay, $\tau_{total}(d_i) = \tau_{prop}(d_i) + \tau_{proc}(d_i) + \tau_{queue}(d_i)$; a feasible assignment must satisfy $\tau_{total}(d_i) \le L_{req}(d_i)$. The associated optimization problem is to choose F so as to minimize the total processing and transfer cost across all data units, subject to this latency constraint and to the capacity bound of each layer.

C. Data Distribution Criteria

Four categories of criteria jointly determine $F(d_i)$. Temporal criteria classify d_i by latency requirement: $L_{req}(d_i) <$ 10 ms mandates edge processing; 10 ms $\le L_{req}(d_i) <$ 100 ms favors the fog layer; $L_{req}(d_i) \ge$ 100 ms permits cloud processing. High-frequency sensor streams (generation rate above 100 Hz) are additionally pre-filtered at the edge to prevent network saturation. Semantic criteria concern whether d_i can be interpreted in isolation or requires aggregation with data from other sources; single-sensor readings are typically resolved locally, whereas area-wide optimization (e.g. city-scale traffic-flow analysis) requires fog- or cloud-level aggregation. Security criteria are based on $S(d_i)$: when $S(d_i) \ge 4$ and the available cloud infrastructure is hosted outside the jurisdiction responsible for the data, d_i must not leave the edge or fog layer, consistent with data-sovereignty requirements such as the EU General Data Protection Regulation [3]. Resource criteria compare the computational demand $C_{req}(d_i)$ with the currently available capacity of each layer; if $C_{req}(d_i)$ exceeds the available edge capacity, d_i is forwarded to the fog or cloud layer, and if the aggregate raw-data rate exceeds link bandwidth, edge-side filtering or compression is required before transfer.

D. Multi-Criteria Decision Function and Dynamic Adaptation

The four criteria above are combined into a single decision through a weighted score assigned to each candidate layer $l \in \{edge, fog, cloud\}$:

$$Score(l) = w_1 \cdot LatencyScore(l, L_{req}) + w_2 \cdot SecurityScore(l, S) + w_3 \cdot ComputeScore(l, C_{req}) + w_4 \cdot LoadScore(l, \rho_l), \quad (6)$$

where $w_1, \ldots, w_4 \ge 0$ are weighting coefficients satisfying $w_1+w_2+w_3+w_4 = 1$, and $\rho_l \in [0, 1]$ denotes the current load density of layer l; the four component scores rate the suitability of layer l with respect to the corresponding criterion. The chosen layer is:

$$Q = \operatorname{argmax}_{l \in \{edge,fog,cloud\}} Score(l). \quad (7)$$

The weights are set according to application priorities: a delay-critical application increases w_1, whereas a privacy-critical application increases w_2.

To account for time-varying network and load conditions, the system state at time t is defined as the triple $S(t) = (\rho(t), B(t), L(t))$, where $\rho(t)$ is the vector of per-layer load densities, $B(t)$ is the vector of available link bandwidths, and $L(t)$ is the vector of current link latencies. A transition policy $\pi: S \to A$, with A the space of admissible layer-assignment actions, selects the destination layer for incoming data; the long-term objective is to maximize the cumulative reward $R = \Sigma_t\, r(t)$, where the instantaneous reward $r(t)$ reflects the degree to which latency requirements are met, load is balanced across layers, and network resources are used efficiently. As a concrete instance of this mechanism, an edge node whose load density exceeds a threshold $\rho_{threshold} = 0.80$ redirects incoming data to a neighboring, less-loaded edge node, or to the fog layer if all nearby edge nodes are similarly saturated.

III. CONCLUSION

This paper has presented a formal model of data distribution in a three-tier Edge–Fog–Cloud architecture. The system was represented as a labeled graph with layer-partitioned nodes and a characterized data set; a layer-assignment function, four categories of distribution criteria, a weighted multi-criteria decision function, and a state-dependent dynamic adaptation mechanism were formalized and linked to the underlying architectural constraints of latency, capacity and data sensitivity. The resulting framework offers a mathematically consistent basis for reasoning about latency-aware and privacy-aware data placement in Edge–Fog–Cloud systems and is intended to underpin the simulation and performance analysis to be carried out in subsequent stages of this research.

REFERENCES

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, “Fog computing and its role in the internet of things,” in Proc. 1st Ed. MCC Workshop on Mobile Cloud Computing (MCC ’12), ACM, New York, NY, USA, 2012, pp. 13–16. doi: 10.1145/2342509.2342513.

[2] IEEE Standards Association, IEEE Std 1934-2018 – IEEE Standard for Adoption of OpenFog Reference Architecture for Fog Computing, IEEE, Piscataway, NJ, USA, 2018. doi: 10.1109/IEEESTD.2018.8423800.

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88.

ФОРМАЛЬНАЯ МОДЕЛЬ РАСПРЕДЕЛЕНИЯ ДАННЫХ НА ОСНОВЕ АРХИТЕКТУРЫ EDGE–FOG–CLOUD

Клычев А.

Инженерно-технологический университет Туркменистана имени Огуз хана, Ашхабад, Туркменистан, annamyrat.gylyjov@etut.edu.tm

Аннотация: системы «умного города» и IoT порождают неоднородные данные, требования которых по задержке, конфиденциальности и вычислительной нагрузке не могут быть эффективно удовлетворены одним уровнем вычислений. В статье формализуется распределение данных в иерархической архитектуре Edge–Fog–Cloud, представленной в виде размеченного графа, узлы которого разделены на слои edge, fog и cloud с различной мощностью и диапазонами задержки. Функция распределения сопоставляет каждой единице данных обрабатывающий слой по её задержке, уровню чувствительности, нагрузке и текущей загрузке ресурсов. Формализованы четыре категории критериев – временные, семантические, безопасности и ресурсные – а также взвешенная многокритериальная функция оценки и механизм динамической адаптации, образующие основу для размещения данных с учётом задержки и конфиденциальности.

Ключевые слова: граничные вычисления; туманные вычисления; облачные вычисления; формальная модель; распределение данных; задержка; интернет вещей; распределённые системы.