CAMP: Cost-Aware Multiparty Session Protocols

DAVID CASTRO-PEREZ and NOBUKO YOSHIDA, Imperial College London, United Kingdom

This paper presents CAMP, a new static performance analysis framework for message-passing concurrent and distributed systems, based on the theory of multiparty session types (MPST). Understanding the runtime performance of concurrent and distributed systems is of great importance for the identification of bottlenecks and optimisation opportunities. In the message-passing setting, these bottlenecks are generally communication overheads and synchronisation times. Despite its importance, reasoning about these intensional properties of software, such as performance, has received little attention, compared to verifying extensional properties, such as correctness. Behavioural protocol specifications based on sessions types capture not only extensional, but also intensional properties of concurrent and distributed systems. CAMP augments MPST with annotations of communication latency and local computation cost, defined as estimated execution times, that we use to extract cost equations from protocol descriptions. CAMP is also extendable to analyse asynchronous communication optimisation built on a recent advance of session type theories. We apply our tool to different existing benchmarks and use cases in the literature with a wide range of communication protocols, implemented in C, MPI-C, Scala, Go, and OCaml. Our benchmarks show that, in most of the cases, we predict an upper-bound on the real execution costs with < 15% error.

CCS Concepts: • Theory of computation → Program analysis; • Software and its engineering → Software performance; System description languages;

Additional Key Words and Phrases: parallel programming, session types, cost models, message optimisations

ACM Reference Format:

1 INTRODUCTION

Understanding the amount of resources, e.g. time or memory that are required by a computation, is of great importance. Correct but slow-performing software can cause a number of problems, ranging from the unnecessary use of resources, to exploitable security vulnerabilities. Worse still, performance issues are very difficult to detect in runtime because of their non fail-stop nature; and although the root causes of performance bugs can be very diverse, uncoordinated functions and synchronisation issues are prevalent, i.e. inefficient composition of efficient functions, and unnecessary synchronisation that increases thread competition [Jin et al. 2012]. These inefficient compositions have more impact in a distributed setting, where the communication overhead and synchronisation cost may become the bottleneck of the whole system.

The development of new static performance analysis tools will reduce the impact of bad performing software, by allowing the identification of their bottlenecks and optimisation. Further, for concurrency and distribution, such a tool must take into account communication and synchronisation overheads. This paper presents a new static performance analysis framework, CAMP.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.
2475-1421/2020/11-ART155
https://doi.org/10.1145/3428223
(Cost-Aware Multiparty Protocols), that can identify potential performance bottlenecks in concurrent and distributed systems. Specifically, CAMP addresses the following two main challenges: non-determinism and practicality. Firstly, the non-deterministic nature of concurrent and distributed systems makes it hard to reason statically about the performance of alternative interleavings of actions in a program trace; and secondly, making the performance analysis practically useful for already existing implementations is not trivial.

CAMP solves the non-determinism issue by building on top of multiparty session types (MPST) [Coppo et al. 2015; Honda et al. 2008]. MPST is a well-established theory that describes not only extensional, but also intensional information about communicating systems. Specifically, MPST captures the communication structure, or protocol among distributed peers. Protocols appear not only in distributed networks but also in parallel multicore programming as patterns or topologies [Goetz et al. 2006; Lea 1997; Rauber and Rünger 2010; Taubenfeld 2006]. MPST uses global types for describing such protocols from a global point of view, and can be used to ensure deadlock-freedom and session fidelity: every send has a matching receive, and every component of the system complies with its part in the global protocol. Built on the MPST theory, CAMP enables the protocol-based performance analysis, giving a precise abstraction as (correct) communication structures of programs. By tying the analysis to a particular protocol specification that is statically enforced on the system, CAMP solves the issue of non-determinism.

On the practical side, since all we require is a global type, CAMP can be readily applied to existing implementations, as long as they are proven to comply with a known global type. We show this by taking existing benchmarks, either implemented using MPST-based tools, or following a known protocol. Different extensions of the core MPST have been already used to implement a wide range of applications written in different programming languages through several transports and architectures e.g. [Castro et al. 2019; Castro-Perez and Yoshida 2020b; Gay and Ravara 2017; Hu and Yoshida 2017; Imai et al. 2020; Ng et al. 2015], and our methodology is easily adaptable to these variants. In addition, not only CAMP is immediately usable for analysis of representative parallel patterns [Asanovic et al. 2009; Krommydas et al. 2016; Rauber and Rünger 2010], but also it is applicable to Savina benchmarks [Imam and Sarkar 2014] or multicore algorithms which incur more complex patterns and synchronisations (§8.2).

The key notion in CAMP is that of execution cost: the amount of time that it takes a protocol, participant or function to run from beginning to end. To statically compute execution costs for concurrency and distribution, CAMP extends global types with sizes for values of messages (encoded in the payload types) and local computation cost information. This size and cost information can be obtained via profiling, or further static analysis, such as using sized-types [Hughes et al. 1996]. Our cost models take these extended protocols, and compute a set of equations which describe the total cost of each participant. These measurements provide us with fine-grained information to obtain communication overhead and synchronisation cost among participants of a protocol. For recursive protocols, CAMP produces a set of recurrence equations that describe the total cost after each iteration of the protocol. For non-terminating protocols (e.g. streaming computation split in multiple stages), CAMP computes the latency, or the average cost per iteration.

CAMP enables to quantify the performance gain of asynchronous communication optimisation. We evaluate this using non-optimised and optimised benchmarks. The optimisation analysis by CAMP is grounded on asynchronous session subtyping, which is one of the most advanced session types theories in the literature, and has been actively studied over a decade using various different formalisms, e.g., first and higher-order mobile processes [Chen et al. 2017, 2014; Ghilezan et al. 2021; Mostrous and Yoshida 2009, 2015; Mostrous et al. 2009], denotational semantics [Demangeon and Yoshida 2015; Dezani-Ciancaglini et al. 2016] and automata theories [Bravetti et al. 2019, 2017, 2018; Lange and Yoshida 2017].
Contributions. We present a compile-time performance analysis framework, CAMP, for concurrent/distributed systems that infers upper execution cost bounds of multiparty session protocols. The cost models in CAMP are parametric, and can combine both static and dynamic (e.g., profiling) information to produce accurate results. We prove that the cost analysis by CAMP is sound with respect to the operational semantics of a given global type instrumented with sizes and execution costs; and extensible to analyse communication optimisations. Our main contributions are:

a) we define the semantics of CAMP, integrating global and local type semantics with local computation costs, that can be used to explore the costs of particular traces (§3);
b) we instrument global types with size and local cost information, and use it to statically estimate an upper bound of the execution cost of a protocol, that we prove sound with respect to the operational semantics (§4);
c) we define multiple metrics on the cost recurrences associated with recursive protocols, that can be be used to effectively analyse the performance behaviour of potentially infinite executions (§5);
d) we extend CAMP to handle asynchronous message optimisations, enabling us to statically quantify the potential performance gains when performing such optimisations/reordering (§6);
e) we implement a DSL for specifying global types, from which we can extract cost equations (§7), and we compare our cost model predictions with real benchmarks used in MPST implementations in different languages: C-MPI [Ng et al. 2015], C-pthreads [Castro-Perez and Yoshida 2020b], Go [Castro et al. 2019], OCaml [Imai et al. 2020] and F★ [Zhou et al. 2020]. Additionally, we apply CAMP to a subset of the Savina benchmarking suite (Scala) [Imam and Sarkar 2014]. These benchmarks include examples of common, and complex topologies, such as ring, butterfly and a double-buffering protocol (§8).

§9 discusses related work and §10 concludes the paper. See Castro-Perez and Yoshida [2020a] for additional definitions and full proofs. The git repository https://github.com/camp-cost/camp provides a working prototype implementation, described in §7 and the data used in §8, with instructions for replicating our experiments.

2 OVERVIEW

MPST Basics. We first explain how MPST satisfies extensional properties. Fig. 1 depicts the standard top-down methodology of MPST enhanced with cost-analysis, which we illustrate by a simple scatter-gather example between two Masters (\(m_1, m_2\)) and two Workers (\(w_1, w_2\)).

\[
G = m_1 \rightarrow w_1 \{r_1\}, m_1 \rightarrow w_2 \{r_1\},
\]

\[
w_1 \rightarrow m_2 \{r_2\}, w_2 \rightarrow m_2 \{r_2\}
\]

\(G\) is a global type: a specification of the protocol between participants from a global perspective. \(G\) says master \(m_1\) first sends a message with type \(\tau_1\) to worker \(w_1\) then to worker \(w_2\), and finally master \(m_2\) collects a message with type \(\tau_2\) from each worker. For each participant \(p\), the global type is projected to a local type, which describes localised send and receive actions from \(p\) viewpoint:

\[
L_1 = w_1 \mid \tau_1, w_2 \mid \tau_1, \text{end} \quad L_2 = m_1 \mid \tau_1, m_2 \mid \tau_2, \text{end} \quad L_3 = w_1 \mid \tau_2, w_2 \mid \tau_2, \text{end}
\]

\(L_1\) says \(m_1\) should send (\(\mid\)) a \(\tau_1\) message to \(w_1\), then to \(w_2\), while \(L_2\) says \(w_1\) receives (\(\mid\)) a \(\tau_1\) message from \(m_1\), followed by sending a \(\tau_2\) message to \(m_2\) (worker \(w_2\) has the same type \(L_2\)). Local types
are used to statically check local programs $P_l$ implementing these types, i.e. the communication structures of each program complies with their local type. A well-typed system of programs is guaranteed free from deadlock and type errors, following the protocol given by $G$ (session fidelity).

**Cost-Aware MPST.** Now we consider the cost to run $G$: $m_1 \mapsto w_1 \{\tau_1 \circ c_1\}. m_1 \mapsto w_2 \{\tau_2 \circ c_1\}. m_1 \mapsto m_2\{\tau_2\}. w_2 \mapsto m_2\{\tau_2\}$ Here $c_1$ represents the local computation cost at the receiver side. In this example, we are assuming the computational cost at $w_1$ and $w_2$ is $c_1$, while such cost at $m_2$ is 0. Another factor we should take into account is the communication cost, which is parameterised by types, i.e. the time required for sending ($c_0(\tau)$) and receiving ($c_1(\tau)$) a value of type $\tau$.

We assume our transport is asynchronous, i.e. sending is non-blocking and the order of messages are preserved, like TCP communications, hence there is no communication ordering between $w_1$ and $w_2$. Both workers can process the values independently at two different locations or CPUs. Then, the total execution cost at $m_1$, $w_1$ and $w_2$ are:

$$ m_1 \mapsto 2 \times c_0(\tau_1) \quad w_1 \mapsto c_0(\tau_1) + c_1(\tau_1) + c_1 + c_0(\tau_2) \quad w_2 \mapsto 2 \times c_0(\tau_1) + c_1(\tau_1) + c_1 + c_0(\tau_2) $$

To consider the cost for $m_2$, we should take care of the dependencies in the protocol. Each $w_i$ can operate in parallel, and they exhibit almost the same cost. The only difference is that worker $w_1$ can perform its computation as soon as $m_1$ sends one message, but worker $w_2$ can only proceed after $m_1$ sends the second message. This difference means that $m_2$ can start gathering one of the messages, while the other worker finishes its actions, which will be delayed by the time it takes $m_1$ to send one message. Hence the cost of $w_2$ is:

$$ m_2 \mapsto \max(c_1(\tau_2), c_0(\tau_1)) + c_0(\tau_1) + c_1(\tau_1) + c_1 + c_0(\tau_2) + c_1(\tau_2) $$

In §4, we shall prove the cost calculated based on local types and global types semantics coincide.

In many scenarios, we do not know how many iterations recursive protocols are going to run, or this number of iterations is large. In such cases, computing the cost of the protocol is not useful or meaningful. In such scenarios, we calculate the average cost per iteration of a protocol (latency) from a set of recurrences. From this latency, we calculate other useful metrics, such as the latency divided by the number of messages exchanged per iteration by participant (latency relative to a particular participant). The latency relative to a participant is used to estimate how much work can a participant do per iteration of the protocol.

### 3 COST-AWARE MULTIPARTY SESSION PROTOCOLS

This section introduces cost-aware multiparty session protocols (CAMP) which is an extension of multiparty session types (MPST) [Coppo et al. 2015; Deniélou and Yoshida 2013; Honda et al. 2008] where the payload types ($\tau$) are types that have been extended with size annotations, adapted from the literature on sized types [Avanzini and Dal Lago 2017; Hughes et al. 1996], and interactions have been extended with cost annotations ($c$) which represent the local execution time at the receiver:

$$ \tau := \text{int}^i | \cdots | D^i r $$
$$ c := i \ | \ k \ | \ e + c \ | \ \text{max}(e, c) \ | \ k \times c $$

Our types are base types (integer, boolean, . . .) annotated with a size $i$, type constructors $D$ applied

Fig. 2. A ring protocol, and an execution trace.
to a sequence of sized types $\overline{\tau}$, annotated with a size $i$. Cost expressions $c$ are either sizes $i$, constants $k$, the addition of two costs, the maximum of two costs, or a constant multiplied by a cost. A size $i$ an arithmetic expression that may contain constants $(k)$ or size variables $(n, m, \ldots)$. Definitions of global and local types are based on the most commonly used MPST in the literature [Coppo et al. 2015]. The syntax of global ($G$) and local ($L$) types in MPST is given below:

$$G \equiv p \to q \{\tau \circ c\}.G \mid p \to q : \{l_i.G_i\}_{i \in I} \mid \mu X.G \mid X \mid \text{end}$$

$$L \equiv p \! \uparrow \tau.L \mid p \oplus \{l_i.L_i\}_{i \in I} \mid p ? \tau \circ c.L \mid p & \{l_i.L_i\}_{i \in I} \mid \mu X.L \mid X \mid \text{end}$$

We start with a set of roles, $p, q, \ldots$, and a set of labels, $l_1, l_2, \ldots$. These are considered as natural numbers: roles are participant identifiers, e.g. thread or process ids; and labels are tags that differentiate branches in the data/control flow. Global type $p \to q \{\tau \circ c\}.G$ denotes data interactions from role $p$ to role $q$ with value of type $\tau$ and local computation cost $c$; Branching is represented by $p \to \{l_i.G_i\}_{i \in I}$ with actions $l_i$ from $p$ to $q$. end represents a termination of the protocol. $\mu X.G$ represents a recursion, which is equivalent to $[\mu X. G/X]G$. We assume recursive types are guarded [Pierce 2002].

Each role in $G$ represents a different participant in a parallel process. Local types represent the communication actions performed by each role. The send type $p \! \uparrow \tau.L$ expresses sending of a value of type $\tau$ to role $p$ followed by interactions specified by $L$. The receive type $p ? \tau \circ c.L$ receives a value of type $\tau$ from role $p$ with local computation cost $c$. The selection type represents the transmission to role $p$ of label $l_i$ chosen in the set of labels $(i \in I)$ followed by $L_i$. The branching type is its dual. The rest are the same as $G$.

Remark 3.1. Global types which combine label and data messages are also used in the literature. They can be encoded as global types in this paper by using singleton labels (see [Deniélou and Yoshida 2013, p.178]). E.g. $p \to q : \{l_i(t_i).G_i\}_{i \in I}$ is encoded as $p \to q : \{l_i.G'_i\}_{i \in I}$ and $G'_i = p \to q \{t_i\}.G_i$. It is possible to account for the differences in cost by setting the cost of sending/receiving labels appropriately, e.g. removing the cost of sending labels, and slightly increasing the size of the data messages, to account for the fact that they must be sent alongside a label.

**End Point Projection.** The local type $L$ of a participant $p$ in a global type $G$ can be obtained by the end point projection (EPP) of $G$ onto $p$, denoted by $G \mid p$. The local type gives a local view of a global protocol onto each participant. Our definition of EPP follows the standard projection rules in [Demangeon and Honda 2012; Deniélou and Yoshida 2013]. The projection uses the full merging operator [Demangeon and Honda 2012; Deniélou and Yoshida 2013], which allows more well-formed global types than the original projection rules [Honda et al. 2008].

**Definition 3.1 (Projection and Merging).** The end point projection (EPP) of $G$ onto $p$, denoted by $G \mid p$, is the partial function defined below, together with the merging of local types $L_i$:

**Projection:**

\begin{align*}
(q \to r \{\tau \circ c\}.G) \mid p &= \begin{cases} 
q \uparrow r.G \mid p & \text{if } p = q \neq r \\
q ? \tau \circ c. (G \mid p) & \text{if } p = q \neq q \\
G \mid p & \text{otherwise}
\end{cases} \\
(q \to r : \{l_i.G_i\}_{i \in I}) \mid p &= \begin{cases} 
r \oplus \{l_i.G_i\} \mid p & \text{if } p = q \neq r \\
q \& \{l_i.G_i\} \mid p & \text{if } p = q \neq q \\
\bigwedge_{i \in I}(G_i \mid p) & \text{otherwise}
\end{cases}
\end{align*}

\begin{align*}
(\mu X.G) \mid p &= \begin{cases} 
\mu X.(G \mid p) & \text{if } G \mid p \neq X', \forall X' \\
\text{end} & \text{otherwise}
\end{cases} \\
(X) \mid p = X & \text{end} \\
\text{end} & \text{if } p = \text{end}
\end{align*}
Merging:

\[ p \& \{ l_1.L_1 \}_{i \in I} \cap p \& \{ l_j.L_j' \}_{j \in J} = p \& \{ l_k.L_k \cap L_k' \}_{k \in I \cap J} \cup \{ l_i.L_i \}_{i \in I \setminus J} \cup \{ l_m.L_m \}_{m \in J \setminus I} \]

\[ \mu X.L_1 \cap \mu X.L_2 = \mu X.(L_1 \cap L_2) \quad L \cap L = L \]

The first line of the projection rule defines a case where the sender and receiver are the same [Deniélou et al. 2012]. The global type projection onto a role is not necessarily defined. Particularly, projecting \( q \rightarrow r : \{ l_i.G_i \}_{i \in I} \) onto \( p \), with \( r \neq p \) and \( q \neq p \), is only defined if the projection of all \( G_i \) onto \( p \) can be merged (Def. 3.1). Two local types can be merged only if they are the same, or if they branch on the same role, and their continuations can be merged. For example, \( p \)'s local type of the global type \( p \rightarrow q \{ \tau \text{oc} \}.\text{end} \) is \( q \) \( r \).end, while \( q \)'s is \( p \) \( \tau \text{oc}.\text{end} \). As a more complex example, \( r \)'s local type of the branching global type:

\[ \mu X.p \rightarrow q \{ l_1.q \rightarrow r : l_2.p \rightarrow r : l_3.X, l_4.q \rightarrow r : l_5.p \rightarrow r : l_6.\text{end} \}

is \( \mu X.q \& \{ l_2.p \& \{ l_3.X \}, l_5.p \& \{ l_6.\text{end} \} \} \).

We say that a global type is well formed, if its projection on all its roles is defined. We denote: \( \text{Wf}(G) = \forall p \in \text{roles}(G), \exists L, G \mid p = L \).

Definition 3.2 (Label Broadcasting). We define a macro to represent the broadcasting of a label to multiple participants in a choice. We write \( p \rightarrow \{ q_j \}_{j \in [1,n]} : \{ l_i.G_i \}_{i \in I} \) as a synonym to \( p \rightarrow q_1 \{ l_1.p \rightarrow q_2 \{ l_1.p \rightarrow \ldots \rightarrow q_n \{ l_i.G_i \ldots \} \}_{i \in I} \}. \) Similarly, for local types, \( \{ q_j \}_{j \in [1,n]} \oplus \{ l_i.L_i \}_{i \in I} \) expands to \( q \oplus \{ l_i.q_2 \oplus \{ l_i.q_1 \oplus \{ l_i.L_i \ldots \} \}_{i \in I} \} \).

It is straightforward to derive: \( p \rightarrow \{ q_j \}_{j \in J} : \{ l_i.G_i \}_{i \in I} \mid r = \{ q_j \}_{j \in J} \oplus \{ l_i.G_i \mid r \}_{i \in I} \), if \( p = r \), and \( p \rightarrow \{ q_j \}_{j \in J} : \{ l_i.G_i \}_{i \in I} \mid r = q_j \oplus \{ l_i.G_i \mid r \}_{i \in I} \), if \( r = q_j \) for some \( j \in J \).

3.1 Labelled Transition System of Global Types

We introduce the labelled transition system (LTS) of global types to associate protocol execution costs with cost annotations. Our semantics is based on the LTS for global and local types in Deniélou and Yoshida [2013] that define their asynchronous operational semantics, and prove their sound and complete correspondence.

We designate the observables \((\alpha, \alpha', \ldots)\) to be the send, receive, branch and select actions that trigger a transition, and an internal transition at a role, which represents the cost \( c \) a role spends performing computation at the receiver (denoted by \( p \circ \text{oc} \)). The syntax of the observables is:

\[ \alpha ::= pq \mid \tau \mid pq \oplus \tau \mid pq \& \tau \mid p \circ \text{oc} \]

The \( \text{poc} \) does not affect the communication structure of the protocol, similar to the silent actions of common process calculi. We say that the subject of an action \( \alpha \) is the role in charge of performing it: \( p = \text{subj}(pq \mid \tau) = \text{subj}(qp \mid \tau) = \text{subj}(pq \oplus) = \text{subj}(pq \&) = \text{subj}(p \circ \text{oc}) \).

Following [Deniélou and Yoshida 2013], we extend the grammar of \( G \) to represent the intermediate steps in the execution with the construct \( p \leadsto q \{ \tau \circ \text{oc} \}.G \) to represent the fact that \( p \) has sent the message of type \( \tau \) but \( q \) has not received it yet, and \( p \circ (\tau \circ \text{oc}).G \) to represent that \( p \) is performing a computation of type \( \tau \) and cost \( c \). For the branching we use \( p \leadsto q \{ l_i.G_i \}_{i \in I} \) to represent the fact that \( p \) has sent label \( l_j \) to \( q \). Then the LTS for global types is defined as below. The main rules different from [Deniélou and Yoshida 2013] are \([GR1a,GR2a,GR2b]\) which consider the execution cost. When we send a message or a label, the type becomes the received mode \( p \leadsto q \) (e.g. \([GR1a]\) and then it asynchronously receives the corresponding message (e.g. \([GR2a]\)). We also observe the actions under the prefix if the participants are unrelated (e.g. \([GR4a]\)).
3.2 Labelled Transition System of Local Types

The labelled transition system (LTS) of local types are given for configurations (C) which map each participant to its local type and a set of FIFO queues (Q) where each represents a queue from sender p to receiver q. We also extend the syntax \((\tau)oc.L\) to represent the intermediate state where the receiver executes a local computation with cost c. In the definition below, [LR2, LR3] formalise the observability of the local computation cost c when receiving the value. Other rules are the standard FIFO enqueue and dequeue rules.

Definition 3.4 (LTS for Local Types). The relation \(\langle C, Q \rangle \overset{\ell}{\rightarrow} \langle C', Q' \rangle\) where \(C = [p \mapsto L_i]_{i \in I}\) and \(Q = [pq \mapsto w]_{i,j \in I,\ell \in I}\) is defined as follows:

\[
\begin{align*}
&[LR1] \quad \langle C[p \mapsto q \! \!; L], Q[pq \mapsto w] \rangle \overset{\text{pe}^{\tau}}{\rightarrow} \langle C[p \mapsto L], Q[pq \mapsto a \cdot w] \rangle \\
&[LR2] \quad \langle C[p \mapsto q \! \!; \tau)oc.L], Q[qp \mapsto w \! \!; \tau] \rangle \overset{\text{pc}^{\tau}}{\rightarrow} \langle C[p \mapsto (\tau)oc.L], Q[qp \mapsto w] \rangle \\
&[LR3] \quad \langle C[p \mapsto (\tau)oc.L], Q \rangle \overset{\text{pc}^\circ}{\rightarrow} \langle C[p \mapsto L], Q \rangle \\
&[LR4] \quad \langle C[p \mapsto q \! \!; \{L_i,L_i\}], Q[pq \mapsto w] \rangle \overset{\text{pq}^{\circ} \! \!; \ell}{\rightarrow} \langle C[p \mapsto L_i], Q[pq \mapsto \ell \cdot w] \rangle \\
&[LR5] \quad \langle C[p \mapsto q \! \!; \{L_i,L_i\}], Q[qp \mapsto w \! \!; \tau] \rangle \overset{\text{pq}^{\circ} \! \!; \ell}{\rightarrow} \langle C[p \mapsto L_i], Q[qp \mapsto \ell \cdot w] \rangle
\end{align*}
\]

The following theorem shows the global type semantics is exactly matched with local asynchronous interactions between participants.

Theorem 3.5. [Soundness and Completeness] Let G be a global type with \(P = \text{roles}(G)\) and let \(C = [p \mapsto G \mid p \in P]\). Then \(G \approx (C; [pq \mapsto e]_{p,q \in P})\).

Proof. (Sketch) The proof of soundness and completeness is a straightforward adaptation of that in [Deniélou and Yoshida 2013]. The distinction between send/receive and select/branch actions is straightforward, e.g. actions GR1a and GR1b are special cases of rule GR1 in [Deniélou and Yoshida 2013] (see Remark 3.1). The addition of cost actions to local types is does not complicate the proof, since it only happens after communication has taken place, which is ensured by being local to each role, and the local context.

Definition 3.6 (Deadlock-freedom). We call \(\langle C_0, Q_0 \rangle\) deadlock-free if for all \(C, Q\) such that \(\langle C_0, Q_0 \rangle \overset{\ell}{\rightarrow} \langle C, Q \rangle\), either (1) \(\forall p \in \text{dom}(C). C(p) = \text{end}\) and \(Q = \emptyset\); or (2) \(\langle C, Q \rangle \overset{\ell}{\rightarrow} \langle C', Q' \rangle\) for some \(C'\) and \(Q'\). We call global type \(G\) deadlock-free if \(\langle C_0, 0 \rangle\) such that \(C_0 = [p \mapsto G \mid p \in P]\) with \(P = \text{roles}(G)\) is deadlock-free.

Note that the definition of deadlock-freedom is not affected by cost annotations. Hence by Remark 3.1, we can directly apply the result in [Deniélou and Yoshida 2013] to obtain:

Theorem 3.7 (Deadlock-freedom). ([Deniélou and Yoshida 2013]) \(\text{Wf}(G)\) is deadlock-free.
4 COST FOR MULTIPARTY SESSION PROTOCOLS (1): BOUNDED RECURSION

This section presents the cost analysis for protocols with bounded recursions. We first define the cost of a trace, as the total cost accumulated by each participant at the end of the execution of a trace. Next we introduce the cost model using global types and show that it provides an upper bound of the cost of any possible trace for each participant.

4.1 Cost of Local Traces

We first explain several assumptions for giving a calculation of cost. Our first assumption in theory is that participants do not share resources with other participants, i.e. they can run on independent CPUs, with no source of contention such as memory or shared cache.

The cost of a trace is the total execution time taken by each participant to run the protocol from start to end. We compute this total execution cost by tracking the dependency from the input to the output (IO-dependency), and by associating each action in a trace (Def. 3.3) with an execution cost. Assuming that every participant has access to their own set of resources (including CPU) implies that every action in the trace (Def. 3.3) will be triggered as early as possible, e.g. send actions will not be arbitrarily delayed by other actions. This also implies that any pair of actions will happen in parallel if swapped freely according to the semantics.

We explain these assumptions with simple examples. Consider the following trace:

\[
p \to r\{\tau\} \cdot q \to r\{\tau'\} \cdot pr!\tau q!\tau' \cdot p \to r\{\tau\} \cdot q \to r\{\tau'\}
\]

According to Def. 3.3, since \(\text{subj}(pr!\tau) = p \neq q = \text{subj}(qr!\tau')\), because the sender is different in each action, this would be another possible trace for the same global type:

\[
p \to r\{\tau\} \cdot q \to r\{\tau'\} \cdot qr!\tau' pr!\tau \cdot p \to r\{\tau\} \cdot q \to r\{\tau'\}
\]

The intuition is that, since \(p\) and \(q\) are running on different CPUs, and their actions are independent, both \(qr!\tau'\) and \(pr!\tau\) can happen in parallel.

We assume that the cost of the message-passing operations depend on the size of the data that is sent and that the costs of sending and receiving messages are known, and that they are functions on the size of the data. Due to the presence of IO-dependencies, we record when the send actions have happened: a participant cannot perform a computation until received a value, and it cannot receive a value until at least the time it took for the sender to finish sending the data has passed.

Cost Environments, Queues and Trace Cost. To record the cost, we use queues that record when the data becomes available. A cost dependency queue is, similarly to \(Q\) in Def. 3.4, a mapping from pairs of participants to queues of execution times, that records when the data in the queue becomes available. We use \(W\) for these cost queues.

The cost of a sequence of actions is defined as a mapping from participants to the total execution time accumulated by each participant, defined as cost environments. It is computed by adding the cost of each individual action to the cost of the participant that performs it, taking into account the cost dependencies recorded by the queue.

We call mappings from participants to total accumulated costs cost environments. If \(c\) is an execution time estimation and \(R\) is a set of participants, then \(T = \{ r \mapsto c \}_{r \in R}\), with the usual
with the maximum of either the time $c'$ or the total accumulated cost by $r$. We write:

$$T[r' \mapsto c] = T[r \mapsto \max(T(r), c') + c]$$

This operation is used to record cost dependencies. Specifically, $T[c' \mapsto r]$ means that $r$ incurs additional cost $c$, after possibly waiting for an action by an external process with total cost $c'$. If the action that took $c$ time depends on an action that took $c'$ time, then the cost of $r$ will be updated with the maximum of either the time $c'$, or the total accumulated cost by $r$. We write:

$$T[r \mapsto c] \text{ for } T[T(r')] \mapsto c' \text{ and } T[r] \mapsto c \text{ for } T[0] \mapsto c.$$
We can see that the cost is the expected one, where the cost includes the initialisation and finalisation wait for \( q \). Since we assume that each participant can run in parallel to the remaining of the participants, the cost of both traces yield the same result:

\[
C(tr_1)([1],[1]) = C(pq \rightarrow c_0(\tau_1), [pq \rightarrow c_0(\tau_1)])) = C(qoc_1 \cdots ([p \rightarrow c_0(\tau_1)]; q \rightarrow c_0(\tau_1) + c_1(\tau_1), []) = \begin{array}{l}
  p \mapsto 2 \times c_0(\tau_1); \quad q \mapsto c_0(\tau_1) + c_1(\tau_1) + c_1 + c_0(\tau_2); \quad r \mapsto 2 \times c_0(\tau_1) + c_1(\tau_1) + c_1 + c_0(\tau_2); \\
  s \mapsto \max(c_1(\tau_2), c_0(\tau_1)) + c_0(\tau_1) + c_1 + c_0(\tau_2) + c_1(\tau_2)
\end{array}
\]

**Example 4.3 (Parallel Pipeline).** We now show the cost of a fragment of the trace that corresponds with \( G = \mu X. p \rightarrow q \{ \tau_1 \circ c_1 \}, q \rightarrow r \{ \tau_2 \circ c_2 \} \). Two possible traces for two iterations of this protocol are as follows:

\[
{\begin{array}{l}
  tr_1 = pq \rightarrow \tau_1 \cdot pq \rightarrow \tau_1 \cdot pq \rightarrow \tau_1 \cdot qoc_1 \cdot qr \rightarrow \tau_2 \cdot qr \rightarrow \tau_1 \cdot qr \rightarrow \tau_2 \cdot \tau_1 \cdot r \circ c_2 \cdot qr \rightarrow \tau_2 \cdot \tau_1 \cdot r \circ c_2 \\
  tr_2 = pq \rightarrow \tau_1 \cdot pq \rightarrow \tau_1 \cdot qoc_1 \cdot qr \rightarrow \tau_2 \cdot qr \rightarrow \tau_2 \cdot r \circ c_2 \cdot pq \rightarrow \tau_1 \cdot pq \rightarrow \tau_1 \cdot qoc_1 \cdot qr \rightarrow \tau_2 \cdot qr \rightarrow \tau_2 \cdot r \circ c_2
\end{array}}
\]

In the first trace, \( p \) sends first two messages to \( q \). Then, \( q \) receives, computes them and sends the result to \( r \). Finally, \( r \) receives the results, and performs their computation with cost \( c_2 \). The second trace, instead, is the repetition of two single iterations of the protocol, where \( p \) sends one message, \( q \) receives, processes and sends the result to \( r \), and \( r \) performs its local computation. Note, however, that since the cost models assume that each participant runs at a different CPU, the costs of both traces is the same. To help readability, we name \( T_p = c_0(\tau_1), T_q = c_1(\tau_1) + c_1 + c_0(\tau_2) \) and \( T_q = c_1(\tau_2) + c_2 \). The trace cost is:

\[
C(tr)([1],[1]) = C(pq \rightarrow c_0(\tau_1), [pq \rightarrow c_0(\tau_1)])) = C(qoc_1 \cdots ([p \rightarrow c_0(\tau_1)]; q \rightarrow c_0(\tau_1) + c_1(\tau_1), []) = \begin{array}{l}
  p \mapsto 2 \times T_p; \quad q \mapsto T_p + T_q + \max(T_p, T_q); \quad r \mapsto T_p + T_q + T_r + \max(T_p, T_q, T_r)
\end{array}
\]

We can see that the cost is the expected one, where the cost includes the initialisation and finalisation of the protocol, where the costs are added, and a pipeline steady state, where the cost is the maximum of the costs of each participant.

**Example 4.4 (Dependency Cycle).** We change slightly the pipeline example, to illustrate what happens to the trace cost when we introduce a dependency cycle in the protocol. The protocol below is a recursive ping-pong, where \( p \) sends to \( q \), and then \( q \) replies to \( p: \mu X. p \rightarrow q \{ \tau_1 \circ c_1 \}, q \rightarrow p \{ \tau_2 \circ c_2 \} \). There is only one possible trace for such protocol, due to the input/output dependencies among \( p \) and \( q \) (see conditions \( p, q \notin subj(\alpha) \) in Def. 3.3, e.g. [GR4a]). The trace and cost in this instance is:

\[
C(tr)([1],[1]) = [p \mapsto 2 \times (T_p + T_q); \quad q \mapsto c_0(\tau_1) + T_p + 2 \times T_q]
\]

Here, \( T_p = c_0(\tau_1) + c_1(\tau_2) + c_2 \) and \( T_q = c_1(\tau_1) + c_1 + c_0(\tau_2) \). Participant \( p \) needs to send \( \tau_1 \), then wait for \( q \) to complete its part of the protocol, and then receive \( \tau_2 \) and process it. Therefore, the cost per iteration is \( T_p + T_q \) in all cases. For participant \( q \), the situation is slightly different. A single iteration of \( q \) only requires it to wait until \( p \) sends \( \tau_1 \), and then perform its part of the protocol. Hence, the cost is \( c_0(\tau_1) + T_q \). However, on the next iteration, \( q \) needs to wait until \( p \) finishes with its actions for the previous iteration. This implies that the cost of a single iteration for \( q \) \((c_0(\tau_1) + T_q)\) is less than the average cost per iteration \((T_p + T_q)\).
4.2 Cost of Global Protocols

We have introduced a way to compute the cost of one trace. This cost is useful to statically analyse the potential execution times of particular executions of a protocol. However, it is in general not feasible to produce all possible traces to analyse the cost of a concurrent/distributed system. Our global type cost addresses this issue, by providing a syntactic method to estimate an upper bound of the execution cost.

The global type cost produces, just like Def. 4.1, a cost environment, with a per-participant estimation. The protocol will complete when all the participants have finished their tasks, and so the overall cost is the maximum of the cost per participant. The global type cost is a function from a global type, an estimation of the number of iterations for the recursive protocols, and an initial cost environment. For proving completeness, we use a dependency queue as an input to the global type cost, that will only be used at intermediate stages of the execution.

Definition 4.5 (Global Type Cost). Let the maximum operation that combines two cost environments compute a per participant maximum \( \max(T, T') = [p \mapsto \max(T(p), T'(p))]_{p \in \text{dom}(T) \cup \text{dom}(T')} \).

We define the function unfold, that unrolls the recursive protocol \( \mu X.G \ k \) k times:

\[
\text{unfold}^{k+1}(X, G) = [\text{unfold}^k(X, G)/X] G \quad \text{unfold}^0(X, G) = \text{end}
\]

Then the global type cost is defined recursively on the structure of global types:

- \( C(p \rightarrow q\{r \in \mathcal{C}\}.G, \tilde{k})(T, W) = C(G, \tilde{k})(T[p \mapsto c_0(r)][q \mapsto c_1(r) + c], W) \)
- \( C(p \rightarrow q\{i_i.G_i\}_{i \in I}, \tilde{k})(T, W) = \max\{C(G_i, \tilde{k})(T[p \mapsto c_0(1)][q \mapsto c_1(1)], W)\}_{i \in I} \)
- \( C(\mu X.G, k \cdot \tilde{k})(T, W) = C(\text{unfold}^k(X, G), \tilde{k})(T, W) \)
- \( C(\text{end})(T, W) = (T, W) \)

For completeness, and for the proofs, we define the cost rules for the extended global types used in the semantics.

- \( C(p \mapsto q\{r \in \mathcal{C}\}.G, \tilde{k})(T, W[pq \mapsto cp \cdot w]) = C(G, \tilde{k})(T[cp[q \mapsto c_1(\tau) + c], W[pq \mapsto w]) \)
- \( C(p \mapsto q j\{i_i.G_i\}_{i \in I}, \tilde{k})(T, W[pq \mapsto cp \cdot w]) = C(G_j, \tilde{k})(T[cp[q \mapsto c_1(1)], W[pq \mapsto w]) \)
- \( C(p \circ (r \circ \mathcal{C}).G, \tilde{k})(T, W) = C(G, \tilde{k})(T[p \mapsto c], W) \)

We write \( C(G, \tilde{k}) \) to represent \( C(G, \tilde{k})([, []) \). Since the dependency queue is only used in the definitions for the intermediate stages of the execution (\( \rightarrow \)), we can write \( C(G, \tilde{k})(T) \). When we compare the output of the cost functions, we refer to the per-participant cost, ignoring the dependency queue: \( (T, W) \leq (T', W'), \forall p \in T, T(p) \leq T'(p) \).

The first rule in Def. 4.5 explains the cost of an interaction from \( p \) to \( q \). Participant \( p \) needs to send a message, and this is what the cost \( T[p \mapsto c_0(\tau)] \) reflects. Participant \( q \) will receive a value from \( p \), and then take \( c \) time performing a computation. Since \( q \) finishes, we add this dependency to the cost: \([p q \mapsto c_1(\tau) + c]\). The cost of a choice is computed similarly, but to produce an upper bound of the cost of all branches, we compute the maximum cost per-participant. The cost of the intermediate stages of the execution of the protocol \( (p \mapsto q) \) requires accessing the information in \( W \), and retrieving when \( p \) completed the send operation. The cost of a computation \( (p \circ (r \circ \mathcal{C})) \) is added to accumulated cost of participant \( p \). The cost of a recursive protocol uses parameter \( k \) to first unroll the recursion, and then compute the cost. We go back to the Examples 4.2, 4.3 and 4.4 and show the computed cost by their global type.

Example 4.6 (Scatter/Gather). We illustrate the global type cost using the scatter/gather protocol:

\[
G = p \rightarrow q\{\tau_1 \in \mathcal{C}\}.p \rightarrow r\{\tau_1 \in \mathcal{C}\}.q \rightarrow s\{\tau_2\}.r \rightarrow s\{\tau_2\}.\text{end}
\]

1The maximum operation is defined even if \( p \) is not in one of the environments: recall that \( T(p) = 0 \) if \( p \notin \text{dom}(T) \)
We generally write $\mathcal{C}(\mathcal{G})$ with $4.12$, and provide a full proof in Castro-Perez and Yoshida [2020a]. In the formalisation, we use $\mathcal{C}(\mathcal{G}, k)(([]))$ produces: 

\[
\mathcal{C}(\mathcal{G}) = \mathcal{C}(\mathcal{G} \cdot \mathcal{G}) = \mathcal{C}(\mathcal{G}) \cdot \mathcal{C}(\mathcal{G})
\]

The final cost produced by the global type predicts the same as the one taking any possible trace.

**Example 4.7 (Parallel Pipeline).** $G = \mu X. \mathcal{G} \cdot \mathcal{G}$. Applying the cost models with $k = 2$, $\mathcal{C}(\mathcal{G}, 2)(([]))$, produces the same cost as the trace cost. Particularly, for any arbitrary $k$, $\mathcal{C}(\mathcal{G}, k)(([]))$ produces:

\[
\mathcal{C}(\mathcal{G}, k)(([])) = \max\{\mathcal{C}(\mathcal{G}, k)(([]), \mathcal{G}, \mathcal{G})\}
\]

We showed in the previous examples that function $\mathcal{C}(\mathcal{G})$ accurately predicts an upper bound of the cost obtained from any trace of the protocol. We formalise this statement below in Theorem 4.12, and provide a full proof in Castro-Perez and Yoshida [2020a]. In the formalisation, we use unfold($\mathcal{G}$, $\mathcal{G}$) to unroll all recursion variables, using the parameters $k$, i.e. unfold($\mathcal{G}$, $\mathcal{G}$) is defined recursively on $\mathcal{G}$, with the only interesting case unfold($\mu X. G \cdot \mathcal{G}$, $\mathcal{G}$) is unfold($\mu X. G \cdot \mathcal{G}$, $\mathcal{G}$). Function unfold($\mathcal{G}$, $\mathcal{G}$) is only defined if there are enough the size of $\mathcal{G}$ that is of the amount of recursion variables in $\mathcal{G}$.

**Definition 4.9 (Well-Formedness of Dependency Queues).** A dependency queue $W$ is well formed with respect to a global type $G$ if it only contains the values required to compute the cost of $G$.

\[
\forall (i \in I), Wf(G_i, W[pq \mapsto w]) \Rightarrow Wf(p \mapsto q \{t \mapsto c \cdot c \cdot w\})
\]

We generally write $Wf(G, (T, W)) = Wf(G, W)$. 

**Lemma 4.10 (Preservation of Wf).** If $Wf(G, (T, W))$ and $G \xrightarrow{\ell} G'$, then $Wf(G', C(\ell)(T, W))$. 

**Proof.** By induction on the structure of $G \xrightarrow{\ell} G'$. See Castro-Perez and Yoshida [2020a].

This lemma states that if $W$ is well-formed with respect to $G$, and $G'$ results from taking a step in $G$, then the queue that results from $C(\ell)(T, W)$ is well formed with respect to $G'$.

**Lemma 4.11 (Cost Preservation).** If $G \xrightarrow{\ell} G'$, then $C(G')(C(\ell)(T, W)) \leq C(G)(T, W)$. 

**Proof.** By induction on the structure of the derivation of $G \xrightarrow{\ell} G'$. 

This is the main lemma, that states that if $G$ transitions to $G'$ with action $\ell$, then, given an initial cost/queue $(T, W)$, the cost of $G'$ on an initial cost after running $\ell$ on $(T, W)$ is less or equal than the cost of $G$ with initial cost $(T, W)$. The reason why this cost is less or equal, rather than equal, is that a branching may take a lower cost path in the protocol.

When such protocols are run in practice, they are aimed at speeding up some computation on a large, potentially unbounded, stream of tasks. Therefore, computing $C(mw(n), k)$ can be computationally very expensive, or impossible if $k$ is unknown. These scenarios are where the average cost per iteration, or latency ($C^{\omega}(G)$) is more useful. The key property of $C^{\omega}(G)$ is that approximates to $C(G, k)/k$ as $k$ grows. In the protocol above, it is clear that $C^{\omega}(mw(n)) > C^{\omega}(mw(m))$ if $n > m$, due to the greater number of interactions that involve $m_1$ and $m_2$ per iteration. However, unless the cost of the extra interactions outweigh the cost of a computation performed by $w_i$, it is preferable to use $mw(n)$ than $mw(m)$, subject to the available resources. This is where the latency relative to a participant is useful, since it provides a better measurement about how fast is $mw(n)$ processing tasks.

We explain the intuition behind $C^{\omega}(G)$ and $C^{\omega}(G)$ using the master-worker protocol, and $C(mw(n), k)$ from §4. For simplicity, we omit branching:

$$mw(n) = \mu X. m_1 \rightarrow w_1\{roe\} \ldots m_1 \rightarrow w_n\{roe\} \ldots w_1 \rightarrow m_2\{roe\}. X$$
To simplify the calculations, we also assume that all \( w_j \) only do their actions after \( m_1 \) has finished sending all tasks, although the cost models in §4 will predict a lower cost for \( w_1 \) than \( w_j \) if \( i < j \), since they only need to wait until their required data has been sent. We start with \( mw(2) \) and \( k = 1 \):

\[
C(mw(2), 1) \leq \begin{bmatrix} m_1 \mapsto 2 \times c_0(r_1); & w_1 \mapsto 2 \times T_1 + c_1(r_1) + c_1 + c_0(r_2); & m_2 \mapsto 2 \times T_1 + T + 2 \times (c_1(r_2) + c_2) \end{bmatrix}
\]

We named the relevant parts as \( T_1 \), \( T \) and \( T_2 \) for readability’s sake. For any arbitrary \( k > 1 \), we compute the cost \( C(mw(2), k) \) as:

\[
C(mw(2), k) \leq \begin{bmatrix} m_1 \mapsto 2 \times T_1 + (k-1) \times 2 \times T_1; & w_1 \mapsto 2 \times T_1 + T + (k-1) \times \max(2 \times T_1, T); & m_2 \mapsto 2 \times T_1 + T + 2 \times T_2 + (k-1) \times \max(2 \times T_1, T, 2 \times T_2) \end{bmatrix}
\]

There are two parts in this cost that can be distinguished, the fixed cost that corresponds to the initial and final stages of the protocol: \([m_1 \mapsto 2 \times T_1; w_1 \mapsto 2 \times T_1 + T; m_2 \mapsto 2 \times T_1 + T + 2 \times T_2]\), and the latency, which is the cost that increases the more iterations we take. In general, for an arbitrary \( n \), the latency is:

\[
C_\omega(mw(n)) = \begin{bmatrix} m_1 \mapsto n \times T_1; & w_1 \mapsto \max(n \times T_1, T); & m_2 \mapsto \max(n \times T_1, T, n \times T_2) \end{bmatrix}
\]

If we keep increasing the number of workers, the latency will indicate a greater cost. However, in this particular protocol what matters is the cost per message interaction of \( m_i \), which are the workers that respectively distribute tasks and collect the results. We use \( C_\omega(mw(n)) = C_\omega(mw(n))/i \), where \( i \) is the number of message exchanged by \( m_2 \) per iteration:

\[
C_\omega_1(mw(n)) = \begin{bmatrix} m_1 \mapsto T_1; & w_1 \mapsto \max(T_1, T/n); & m_2 \mapsto \max(T_1, T/n, T_2) \end{bmatrix}
\]

Since \( C_\omega_1(mw(n)) \) is less than \( C_\omega_2(mw(m)) \) if \( m < n \), then the latency relative to \( m_2 \) is a better measurement to compare how fast a protocol processes tasks. In the remainder of this section, we define \( C_\omega \) and \( C_\rho \); and prove that they approximate \( C(G, k) \) for a sufficiently large \( k \).

### 5.1 Latency of Nested Recursive Protocols

The master-worker protocol above contains only one recursion variable. In general, recursive protocols can have multiple nested recursive sub-protocols. Intuitively, to compute \( C_\omega \) of \( \mu X.G \), we need to estimate the total execution time of a single iteration of \( G \). If \( G \) contains recursive sub-protocols, this implies that we need to know how many iterations they will run, before recursive variable \( X \) is found. We illustrate this with the recursive global type below:

\[
G = \mu X. \ p \rightarrow q(\tau \omega_1). \mu Y. \ q \rightarrow p \{l_1. Y; l_2. X\}
\]

To compute \( C_\omega(G) \), we need to know how many times the branch that ends in recursion variable \( Y \) will be taken. Since this depends on the particular implementation of the protocol, we parameterise such recursion variables with some \( k \), and defer its instantiation. To produce an equation to estimate the latency that is parametric in this \( k \), we split the protocol \( G \) into two sub-protocols:

\[
G_Y = \mu Y. \ q \rightarrow p \{l_1. Y; l_2. Z\} \quad G_X = \mu X. \ p \rightarrow q(\tau \omega_1).\{X/Z\}G_Y
\]

If \( C_\omega(G_Y) \) contains another recursion variable, then we keep splitting it until we have a set of global types, each of which defined using at most one bound recursion variable. To compute \( C_\omega(G_X) \) we require a parameter \( k \), and we will use \( k \times C_\omega(\{\text{end}/Z\}G_Y) \) for the cost of any participant in the inner sub-protocol. In the remainder of this section, we focus on recursive protocols with at most one recursion variable.
5.2 Cost Recurrences

Computing $C^\omega$ is done in two steps. First, we build a system of recurrence equations, $T(n)$, that capture the execution costs after $n$ iterations of the recursive protocol. Then, we build the difference equations $\Delta(n)$, where $\Delta(n)(p) = T(n+1)(p) - T(n)(p)$, and estimate the value of $\Delta(n)$, as $n$ grows. We observe that, for the recurrences that we generate, with $n \geq 2$, $\Delta(n)$ stabilises.

Definition 5.1 (Cost Recurrences). We use $C(G)$ from Def. 4.5. Given a recursive global type $\mu X.G$, we define its cost recurrence, $T(n)$, as follows: $T(n+1) = C([\text{end}/X]G, T(n))$, $T(0) = \emptyset$.

Consider the following parallel pipeline:

$$G = \mu Y. p \rightarrow q\{\tau_1 \circ c_1\}, q \rightarrow r\{\tau_2 \circ c_2\}. Y$$

We show below an example of the system of recurrence equations that we generate. We take the resulting cost environment, and we produce a different equation $T_p, T_q, T_r$ for every participant in the protocol:

$$T_p(n+1) = T_p(n) + c_0(\tau_1), T_q(n+1) = \max(T_p(n+1), T_q(n)) + c_1(\tau_1) + c_1 + c_0(\tau_2), T_r(n+1) = \max(T_q(n+1), T_r(n)) + c_1(\tau_2) + c_2$$

Definition 5.2 (Cost Difference Equations). Given a recursive global type $\mu X.G$, with cost recurrence $T$, we define its cost difference equation $\Delta$, as $\Delta(n) = T(n+1) - T(n)$.

The cost difference equation provides an estimate on how much the cost increases for each participant after running the protocol one additional iteration.

Definition 5.3 (Latency per Iteration). The latency of a recursive protocol $\mu X.G$ with cost difference $\Delta(n)$ is defined as the cost expression $c$ that is the least upper bound of the difference equation $\Delta(n)$, for a sufficiently large $n$:

$$C^\omega(\mu X.G) = c \quad \text{s.t. } \exists k, \forall n \geq k, c \geq \Delta(n)$$

Suppose that we want to compute the latency of the previous parallel pipeline. On average, excluding the initialisation of the protocol, the latency for $r$ must be the maximum of the times for $p$, $q$ and $r$, as usual in parallel pipelines. This is because the actions of $p$, $q$ and $r$ are independent across iterations. The solution of $\Delta(0)$ shows that the cost is the addition of all individual costs. However, by solving $\Delta(1)$, we obtain the expected result, where $T_p = c_0(\tau_1)$, $T_q = \max(T_p, c_1(\tau_1) + c_1 + c_0(\tau_2))$, and $T_r = \max(T_q, c_1(\tau_2) + c_2)$.

When the actions of a recursive protocol are not independent across iterations, i.e. the send/receive dependency graph forms a cycle, then all participants will need to synchronise. An example of this is the protocol:

$$G = \mu Y. p \rightarrow q\{\tau_1 \circ c_1\}, q \rightarrow p\{\tau_2 \circ c_2\}. Y$$

In the first iteration, we will have that $p$ sends $\tau_1$ to $q$, which needs to wait for the message, and then takes $c_1$ time. At this point, we have that $p$ spent $c_0(\tau_1)$, and $q$ took $c_0(\tau_1) + c_1(\tau_1) + c_1$. Next, $q$ sends $\tau_2$ to $p$, which is completed after $c_0(\tau_1) + T_q$, where $T_q = c_1(\tau_1) + c_1 + c_0(\tau_2)$. Then, $p$ needs to receive $\tau_2$ and take $c_2$ of local computation time. Since the accumulated time by $p$ is $c_0(\tau_1) < c_0(\tau_1) + T_q$, we increase the total time spent by $p$: $T_p + T_q$, where $T_p = c_0(\tau_1) + c_1(\tau_2) + c_2$. In the next iteration, we have that $p$ takes $T_p + T_q + c_0(\tau_1)$. Next, $q$ takes $\max(c_0(\tau_1) + T_q, T_p + T_q + c_0(\tau_1)) = T_q = T_p + T_q + c_0(\tau_1) + T_q$, and finally $p$ will take $2 \times (T_p + T_q)$. After $k$ iterations, the cost for $p$ is $k \times (T_p + T_q)$, while the cost for $q$ is $(k-1) \times (T_p + T_q) + c_0(\tau_1) + T_q$, which approximates $T_p + T_q$.

Definition 5.4 (Latency with respect to $p$). We define $C^\omega_p(\mu X.G) = C^\omega(\mu X.G)(p)/\text{count}(p, G)$, where $\text{count}(p, G)$ is the number of interactions of $G$ in which $p$ occurs.
5.3 Correctness

We guarantee that the latency correctly approximates the bounded cost of a protocol. Moreover, given an arbitrary trace that is the result of a $k$-unrolling of a recursive global type $G, k \times C^{\omega}(G)$ will approximate the cost of the full trace.

**Theorem 5.5** (Cost Latency Correspondence). Given a sufficiently large $k_2$, for all $k_1 > k_2$, $C(G, k_1) - C(G, k_2) \leq (k_1 - k_2) \times C^{\omega}(G)$.

This result follows directly from our definition of $C^{\omega}$, since the latency approximates $\Delta(n)$ (with $\Delta(n) = T(n + 1) - T(n)$) for a sufficiently large $n$, and that $T(n + 1)$ is the recurrence that approximates $C(G, n)$. We need to show that $k \times \Delta(n) = T(n + k) - T(n)$, and then take $k_2 = n$ and $k_1 = n + k$.

**Proposition 5.6.** Given $\mu X. G$, let $T(n+1) = C([\text{end}/X] G, T(n))$ and $T(0) = \emptyset$. Then, $C(\mu X. G, n) = T(n)$.

**Proof.** By induction on $n$, the base case is straightforward: $T(0) = C(\mu X. G, 0) = C(\text{end}) = \emptyset$. If $n = m+1$, then $T(m+1) = C([\text{end}/X] G, T(m)) = C([\text{end}/X] G, C(\mu X. G, m)) = C(\mu X. G, m+1)$.

Proposition 5.6 states that given a recursive protocol, instantiating its recurrence with some number $n$ yields the same cost as unrolling the protocol $n$ times and computing its cost. We use Proposition 5.6 in combination with Definition 5.2 to derive the following. Assume $\Delta$ is the difference equation for recursive protocol $G$. Then, the following equality holds:

$$\Delta(n) = C(G, n + 1) - C(G, n)$$

(1)

**Theorem 5.7** (Latency Soundness). There exists $k'$ such that for all $k$, if unfold$^k(G) \rightarrow \text{end}$, then $C(\tilde{a}) \leq k \times C^{\omega}(G) + k'$.

**Proof.** By Definition 5.2, we know that there exists some $k_0$ such that for all $n \geq k_0$,

$$C^{\omega}(G) \geq \Delta(n).$$

(2)

We show that $k'$ is $C(G, k_0)$. By Theorem 4.12, we know that $C(\tilde{a}) \leq C(G, k)$. Therefore, it is sufficient to show that for all $k$, $C(G, k) \leq k \times C^{\omega}(G) + C(G, k_0)$. We proceed by case analysis:

**Case** $k \leq k_0$ straightforward, since $C(G, k) \leq C(G, k_0)$ if $k \leq k_0$.

**Case** $k > k_0$: By induction on $k$. All cases $\leq k_0$ are straightforwardly true.

- Case $k = k_0 + 1$ follows from $C(G, k_0) \leq C(G, k_0 + 1)$:

$$C(G, k_0) \leq C(G, k_0 + 1)$$

{multiply $k_0$}

$$k_0 \times C(G, k_0) \leq k_0 \times C(G, k_0 + 1)$$

{add $C(G, k_0 + 1)$}

$$C(G, k_0 + 1) \times C(G, k_0) \leq C(G, k_0 + 1) + k_0 \times C(G, k_0 + 1)$$

{sub $k_0 \times C(G, k_0)$}

$$C(G, k_0 + 1) \leq (k_0 + 1) \times C(G, k_0) + (C(G, k_0) - C(G, k_0))$$

{cancel $C(G, k_0)$}

$$C(G, k_0 + 1) \leq (k_0 + 1) \times C^{\omega}(G) + C(G, k_0)$$

{by (1) and (2)}

- $k = k_2 + 1$, with $k_2 > k_0$. Assume the induction hypothesis $C(G, k_2) \leq k_2 \times \Delta(k_2) + C(G, k_0)$:

$$C(G, k_2) \leq k_2 \times \Delta(k_2) + C(G, k_0)$$

{by (1)}

$$C(G, k_2) \leq k_2 \times C(G, k_2 + 1) - k_2 \times C(G, k_2) + C(G, k_0)$$

{add $C(G, k_2 + 1)$}

$$C(G, k_2 + 1) \leq (k_2 + 1) \times C(G, k_2 + 1) - k_2 \times C(G, k_2) + C(G, k_0)$$

{sub $C(G, k_2)$}

$$C(G, k_2 + 1) \leq (k_2 + 1) \times C(G, k_2 + 1) - C(G, k_2) + C(G, k_0)$$

{by (1) and (2)}

$$C(G, k_2 + 1) \leq (k_2 + 1) \times C^{\omega}(G) + C(G, k_0)$$

□

This implies that the latency approximates the cost of a trace of a \( k \)-unrolling of a recursive protocol, and it follows from Theorems 5.5 and 4.12. To illustrate this, consider the average cost per recursion iteration, \( C(\bar{a})/k \). By Theorem 5.7, we know that \( C(\bar{a})/k \leq C^\omega(G) + k'/k \). Since \( k' \) does not depend on \( k \), for a sufficiently large \( k \), the term \( k'/k \) will become smaller, and the upper bound of \( C(\bar{a})/k \) will be approximately \( C^\omega(G) \).

### 6 ASYNCHRONOUS MESSAGE OPTIMISATION

This section illustrates one of the key features of CAMP, the formulation and its soundness of asynchronous message optimisations. We extend the cost equations in §4 and §5 to tackle protocols in which certain actions have been permuted for optimisation purposes. Parallel programs often make use of parallel pipelines to overlap computation and communication, as far as the overlapping does not interfere with data dependencies. The overlapping can reduce stall time due to blocking wait in the asynchronous communication model. Under the CAMP theory, optimisation should preserve the deadlock-freedom and produce the same outcome, while ensuring less cost for the overall calculation.

Fig. 5 shows a safe and efficient ring protocol, in which stage \( i \) shares data with stage \( (i + 1) \mod 3 \), and then proceed to do some local computation. This protocol behaves similarly to that of Fig. 2 in §3, but the output actions have been permuted so that they are performed first, thus reducing the amount of synchronisation required. The optimised version, however, is more difficult to check against a standard global type, because of the permuted actions. This can be illustrated by comparing the optimised and un-optimised local types of \( q \):

\[
L_q = \mu X. ? \tau \cdot c \cdot X \\
L'_q = \mu X. ! \tau \cdot p \cdot ? \tau \cdot c \cdot X
\]

\( L_q \) is the unoptimised local type, and \( L'_q \) is the optimised version. Both local types represent a similar communication pattern. However, in the left version \( L_q \), the send action only happens after receiving, and computing (with cost \( c \)), while the right version first sends a value of type \( \tau \), and then performs the receive and local computation. This removes unnecessary synchronisation, and allows \( r \) to continue with its interactions before \( q \) finishes its own local computation.

Only certain message permutations are valid. For example, if instead of swapping the send and receive actions for \( L_q \), we permute the actions for participant \( p \), then we end up in the following (incorrect) situation:

\[
L'_p = \mu X. r \cdot ? \tau \cdot c \cdot q \cdot X \\
L'_q = \mu X. p \cdot ? \tau \cdot c \cdot r \cdot X \\
L'_r = \mu X. p \cdot ? \tau \cdot c \cdot p \cdot X
\]

This is a clear deadlock situation, since all participants are waiting for a message from each other. To avoid such situations, we define the Asynchronous Message Optimisation for global types, and show its soundness:

**Definition 6.1 (Asynchronous Message Optimisation).** We first extend the syntax of global types to include send (!) and receive (?) actions as: \( G ::= pq ! \tau \cdot G | pq ? \tau \cdot G | \ldots \). The asynchronous optimisation relation, \( G_1 \leq G_2 \) (read: \( G_1 \) is more optimal than \( G_2 \)), with \( p_1 \neq p_2 \) or \( q_1 \neq q_2 \) is the transitive closure of the rules below:
we do not unroll recursion; and b) the only potential rewrite cycle appears in rules 155:18 David Castro-Perez and Nobuko Yoshida deadlock free, then 

\[ \text{Out} \]  

permutes two outputs to two different participants; [In] is its dual; [OBr] permutes a send and a branch; [IBra] is dual; and [Cong] is a congruence rule. The key rules are [Opt], [OBr] and [IBra], that perform permutations that allow communication and computation to overlap. This is because the rules permute send actions to the left, and receive actions to the right. We prove that whenever \( G_2 \) is deadlock-free, then \( G_1 \leq G_2 \) must also be deadlock free. Moreover, we show that \( G_1 \leq G_2 \) is decidable. Notice that: (1) our definition is different from the literature asynchronous subtyping for session types, motivated from more practical use cases; (2) our cost models can be applied even whenever we do not have that \( G_1 \leq G_2 \), in which case, safety can be guaranteed by using any method from the literature. See §9.

**Theorem 6.2 (Asynchronous Message Optimisation).**

1. (Soundness) Suppose \( G_2 \) is a deadlock-free global type and \( G_1 \leq G_2 \). Then \( G_1 \) is deadlock-free.
2. (Decidability) Given \( G_1 \) and \( G_2 \), it is decidable whether \( G_1 \leq G_2 \) or not.

**Proof.** (1) By induction on the derivation of \( G_1 \leq G_2 \). Assume \( G_1 \leq G_2 \) and \( G_2 \) is deadlock-free and \( C_i = [p \mapsto G_i \mid p \in \mathcal{P}] \) with \( i = 1, 2 \). We prove if \( \langle C_2, Q \rangle \) is deadlock-free, then \( \langle C_1, Q \rangle \) is deadlock-free. To do this proof, we extend the projection for global types as follows. \( \text{pq}! \langle \tau \rangle . G \mid r = \text{q}! \langle \tau \rangle . (G \mid r) \) if \( p = r \), and \( G \mid r \) otherwise. \( \text{pq}? \langle \tau \rangle . G \mid r = \text{q}? \langle \tau \rangle . (G \mid r) \) if \( p = r \), and \( G \mid r \) otherwise. All cases except [Opt] is obvious. The [Opt] states: \( p_1 \mid q_1 \mid \{ \tau_1 \} . p_2 \mid \{ \tau_2 \} . G \leq \{ q_2 \} . \{ \tau_2 \} . p_1 \mid \{ \tau_1 \} . G \). We know that \( \langle C_2, Q \rangle \) is deadlock free. Note that \( p_1 \neq q_1 \), otherwise \( G_2 \) cannot be proven deadlock free (it is either ill-formed, or the optimisation of an ill-formed global type). There are two cases, considering the side conditions for the rules: a) if \( p_1 \neq q_2 \), straightforward since these subject of both interactions are different; b) if \( p_1 = q_2 = p \), then we have \( C_1 \setminus p = C_2 \setminus p, C_2(p) = p_2 \mid \{ \tau_2 \} . q_1 \mid \{ \tau_1 \} . L, \) and \( C_1(p) = q_1 \mid \{ \tau_1 \} . p_2 \mid \{ \tau_2 \} . L \). Since \( \langle C_2, Q \rangle \) is deadlock free, then \( Q(p_2, p) = w \cdot \tau_2 \). Therefore, \( \langle C_1, Q \rangle \xrightarrow{\ast} \langle C'_1, Q' \rangle, \langle C_2, Q \rangle \xrightarrow{\ast} \langle C'_2, Q' \rangle, \) with \( C'_1(p) = C'_2(p) = L, \) and \( Q'(p_2, p) = w \). Since \( \langle C_2, Q \rangle \) is deadlock free, then \( \langle C'_2, Q' \rangle \) must also be deadlock free, and \( \langle C'_1, Q' \rangle \) as well.

(2) We consider a normal form which is derived applying [Out,In] with the side condition \( p_1 < p_2 \); and all other rules except [Init] as much as possible until no rule is applicable, and finally applying [Init] to all pairs of send/receive. Then if \( G_1 \leq G_2 \), there exists a unique global type \( G \) such that \( G_1 \leq G \) derivable applying the above rules finitely. This means interpreting \( G_1 \leq G_2 \) as a term rewriting system. The term rewriting system is terminating because a) the terms are finite, since we do not unroll recursion; and b) the only potential rewrite cycle appears in rules [Out,In], which is prevented by the additional side condition that \( p_1 < p_2 \). The repeated application of these rules permute the send and receive actions to their rightmost and leftmost position respectively. By the side conditions of the rules, it is straightforward to show that any critical pairs can be unified, since no rule can prevent another rule from being applied. \( \square \)
Finally, we prove that, ignoring sending costs, if $G_1 \preceq G_2$, then the cost of $G_1$ is less than the cost of $G_2$. The reason why we need to ignore sending costs for this proof is that permuting two output actions may introduce delays in a later computation stage. Note that this property is a statement about the synchronisation costs, not an algorithm for optimising a protocol. To illustrate this case, consider the following global types:

\[
G_1 = \text{pq}! \cdot r \cdot \text{pq}2 \cdot r \cdot \text{q} \cdot p \cdot r \cdot \text{q} \cdot p \cdot r \cdot \text{q} \cdot c_1 \cdot q \cdot c_2 \cdot q \cdot r \cdot \text{rq}2 \cdot r' \cdot \text{rq}2 \cdot r' \cdot \text{end}
\]

\[
G_2 = \text{pq}2 \cdot r \cdot \text{pq}1 \cdot r \cdot \text{q} \cdot p \cdot r \cdot \text{q} \cdot p \cdot r \cdot \text{q} \cdot c_1 \cdot q \cdot c_2 \cdot q \cdot r \cdot \text{rq}2 \cdot r' \cdot \text{rq}2 \cdot r' \cdot \text{end}
\]

It is clear that $G_1 \preceq G_2$, by [Out]. However, whenever $c_2 \geq c_1$, then $C(G_1) \geq C(G_2)$, since $q_2$ must wait longer in $G_1$ than in $G_2$ before receiving the message of type $r$. Note that, even if $c_1 = c_2$, the cost of the global protocol will be greater in $G_1$, since $r$ is the participant that takes longer in the protocol, and needs to wait for $q_2$. The implications of this result are twofold: (a) we know that whenever $G_1 \preceq G_2$, $G_1$ contains less overhead due to synchronisation; and (b) for a given $G_2$, choosing an optimal $G_1 \preceq G_2$ is not straightforward, and depends on actual local computation costs and communication latencies.

**Theorem 6.3 (Optimisation Cost).** Suppose $G_2$ is a well-formed global type and $G_1 \preceq G_2$. If the sending cost is 0, then $C(G_1) \leq C(G_2)$.

**Proof.** By induction on the derivation of $G_1 \preceq G_2$. Most cases are permutations of independent interactions, and all independent interactions can be permuted with no effect on the cost. Since we assume zero send costs, the cost of sending two actions is the same, independently of the order. The reasoning is similar for receiving interactions. The only rules that we need to consider are [Opt], [OBr] and [IBr]. Notice that in all the cases, the left hand side contains a sending (or choice) at an earlier position than the right hand side. We show the proof for case [Opt], but all cases follow a similar structure. The cost of $p_1 \cdot q_1 \cdot \{r_1\} \cdot p_2 \cdot q_2 \cdot \{r_2\} \cdot G$ is the cost of $G$, where the message queue for $p_1 \cdot q_1$ contains the current execution time for $p_1$. If $p_1 \neq p_2$, then the cost will be the same in both cases. But if $p_1 = p_2$, then the cost in the right hand side will contain the accumulated cost for $p_1$, plus the cost of receiving from $q_2$. Since the costs recorded at the message queue are greater, then the cost of the continuation must also be greater. □

We illustrate how this optimisation reduces synchronisation time with one iteration of a ring protocol of size 2: $p \rightarrow q \cdot \{r \cdot c_1\} \cdot q \rightarrow p \cdot \{r_2 \cdot c_2\} \cdot \text{end}$. The only possible trace for running such protocol is: $pq \cdot r_1 \cdot pq ? \cdot r_1 \cdot q \cdot c_1 \cdot qp \cdot r_2 \cdot qp ? \cdot r_2 \cdot p \cdot c_2$. This trace and the derived cost imply that computation costs $c_1$ and $c_2$ cannot happen in parallel: $T_p = c_0(r_1) + c_1(r_1) + c_1 + c_0(r_2) + c_1(r_2) + c_2$. In cases where such interactions are independent, we can permute the send/receive actions of $q$ to remove the synchronisation cost from $p$, and allow any trace that is an interleaving of the following sub-traces, where the send operations happen before the matching receive:

\[
tr_p = pq ! \cdot r_1 \cdot qp ? \cdot r_2 \cdot p \cdot c_2 \quad tr_q = qp ! \cdot r_2 \cdot pq ? \cdot r_1 \cdot q \cdot c_1
\]

Such optimisations are represented by the following type: $pq ! \cdot \{r_1\} \cdot qp ! \cdot \{r_2\} \cdot qp ? \cdot \{r_1 \cdot c_1\} \cdot pq ? \cdot \{r_2 \cdot c_2\} \cdot \text{end}$. This scenario will have the cost that we show below, which is smaller than the original cost.

\[
T_p = \max(c_0(r_1), c_0(r_2)) + c_1(r_2) + c_2 \quad T_q = \max(c_0(r_1), c_0(r_2)) + c_1(r_1) + c_1
\]

## 7 Implementation

We implemented a library in Haskell for describing global types augmented with size and cost information, from which we can derive cost equations for protocols.
7.1 Resource Contention

CAMP addresses the issue that multiple participants may need to share computational resources. We model the cases in which the participants of a protocol are mapped to distinct nodes of a distributed system, where each node may contain multiple cores. This requires: a) a target hardware specification, and b) a mapping from participants to nodes. The target hardware specification describes the amount of nodes available, the cores per-node, and the communication latencies between nodes. The mapping from participants to nodes assigns each participant of the distributed system to a different node. Our assumptions are: a) there is no mechanism for process migration; b) processes can be pinned to specific nodes, but not to specific cores; and c) an optimistic scheduling scenario, in which participants will run as soon as possible, whenever a core becomes available.

Definition 7.1 (Target Hardware Specification). The target hardware is specified as an indexed set of node descriptions, and the communication latencies between nodes: \( \{C_n\}_{n \in \mathbb{N}} \) and \( \{L_{n_1,n_2}\}_{n_1,n_2 \in \mathbb{N}} \). Here, \( \mathbb{N} \) is the set of node identifiers, \( C_n \) is a natural number that describes the number of available cores for node \( n \), and \( L_{n_1,n_2} \) is a function from a size to the amount of time it takes to transmit a value from \( n_1 \) to \( n_2 \).

Definition 7.2 (Participant Mapping). The participant mapping associates each participant with a specific node. We say that participants are pinned to nodes \( M : \mathcal{P} \rightarrow \mathbb{N} \).

For example, consider the master-worker example, where we have 1 master and 5 workers:

\[
\mu X. m \mapsto w_1(\tau_1). \ldots . m \rightarrow w_5(\tau_1). w_1 \rightarrow m(\tau_2). \ldots . w_5 \rightarrow m(\tau_2). X
\]

First, we need to know which is the target hardware. In our case, this is a distributed system with two nodes, \( n_1 \) and \( n_2 \), with 1 and 4 cores respectively. That is: with \( C_{n_1} = 1 \) and \( C_{n_2} = 4 \). Suppose that the communication latency between \( n_1 \) and \( n_2 \) is a known function on the size of the messages, \( l \). Then, \( L_{n_1,n_2} = L_{n_2,n_1} = l \). Our hardware description is completed by \( \{C_n\}_{n \in \{n_1,n_2\}} \), and \( \{L_{n,n'}\}_{n,n' \in \{n_1,n_2\}} \). Finally, we require to map our participants to the different nodes in the architecture. In our example, we may want to run \( m \) in \( n_1 \), and \( w_i \) in \( n_2 \): \( M(m) = n_1 \) and \( M(w_i) = n_2 \).

To compute the cost in this specific scenario, we use the resource bounded cost equations. The key difference is that, as well as keeping track of the accumulated time per-role, we keep the accumulated time per node, using a core-availability time, which is the earliest time at which a core becomes available. The resource-bounded cost equations are obtained using \( C(G)(T, S, W) \), where \( S \) accumulates the cost at each core and each node of the system. We assume a hardware specification and mapping. The rules are now modified in the following way:

\[
C(p \rightarrow q\{\tau\}, G, \tilde{G})(T, S, W) = C(G, \tilde{G})(T)p \mapsto S_1(M(p)_{c_1})q \mapsto S_2(M(q)_{c_2}), S_2, W
\]

where \( S_1 = S[M(p)_{c_1} \mapsto c_0(\tau)] \), \( S_2 = S[M(q)_{c_2} \mapsto c_1(\tau) + L_{M(p)M(q)}] \), \( \forall c, S[M(p)_{c_1}] \leq S[M(p)_{c}] \) and \( \forall c, S_1[M(q)_{c_2}] \leq S[M(q)_{c}] \). In this definition, we update the accumulated cost of \( p \) and \( q \) to the total accumulated cost of the lowest cost core of the node to which they are mapped. The definition of \( S[n_{c} \mapsto c] \) is the same as in §4.

7.2 A Monadic Interface for Global Types

We develop a deep embedding of the global types of §3 in Haskell, and provide a monadic interface on top as a simpler interface for representing protocols. We call this monadic interface GTM, for Global Type Monad. In GTM, there is an implicit end at the end of each sequence of interactions. An interaction is specified using function message, and participants are created using mkRole. Function gclos runs the code in the GTM monad, and produces the resulting global type (CGT). We show below the Haskell code that generates an \( n \)-stage pipeline, and a recursive 2-stage pipeline generated using the following code:
pipe :: [(SType, Cost)] → Role → GTM () → GTM ()
pipe [] p k = k
pipe ((t, c) : r) p k = mkRole >>= λq → message p q t (cost t) >>= pipe r q k

rpipe2 :: CGT
rpipe2 = gclose $ mkRole >>= = λr → grec $ λx → pipe [(t1, c1), (t2, c2)] r x

The code for rpipe2 produces the following global type: μX.p → q{t1⋄c1}.q → r{t2⋄c2}.X. Notice that embedding a global type language in Haskell allows us to compute topologies based on any input parameters, such as the number of stages of a pipeline, that would otherwise require the use of extensions to MPST, e.g. parameterised roles [Castro et al. 2019; Deniélou et al. 2012]. However, to check well-formedness, we need to instantiate the parameters.

We provide functions cost and latency, both of type CGT → Time, to compute the set of equations that describe the cost (latency) of an input global type. To obtain a particular prediction, the user needs to provide an instantiation of all free size and cost variables in the equations, including the transmission costs between participants.

8 EVALUATION

This section presents a number of benchmarks used to evaluate the predictive power of CAMP. Our benchmarks are taken from multiple different sources, mostly MPST-based tools [Castro et al. 2019; Castro-Perez and Yoshida 2020b; Imai et al. 2020; Ng et al. 2015; Zhou et al. 2020], but also a subset of the Savina actor benchmarking suite [Imam and Sarkar 2014]. We categorise our benchmarks following the structure of the Savina benchmarking suite: (i) microbenchmarks, (ii) concurrency benchmarks, and (iii) parallel algorithms. Microbenchmarks focus on different structures and protocols, and are aimed at testing and evaluating the different features of CAMP. Concurrency benchmarks are aimed at evaluating the impact on communication and synchronisation. This can be useful to, e.g. estimate server response times, and set the appropriate timeouts in larger systems. In the context of parallel algorithms, the main use of the cost models is to predict the parallel speedups achieved by a particular parallelisation, without needing to run or profile the application.

8.1 Methodology

We follow a series of steps in order to make our results as consistent as possible. We will detail now these steps, highlighting which part is automated, and which needs to be provided by the developer. Our methodology is divided in two parts: (1) characterising the target architecture; and (2) benchmark cost analysis.

Characterising the target architecture. To tailor a cost analysis to a specific target architecture, we need to characterise the costs of sending/receiving data between nodes. This requires three steps: (1) specifying the amount of nodes, and the amount of processors/cores per node; (2) estimate message latencies between nodes; and, (3) profiling send/receive operations in the required languages/frameworks with inputs of different sizes.

We require the results of these steps to be stored in a .hs file, as an architecture description, that will be imported and used by CAMP’s cost models. These steps must be performed only once per architecture and programming language.

Additionally to our theory, the implementation allows programmers to specify an overhead for running multiple participants in a single node. This is to account for all factors that CAMP is currently not considering for deriving cost equations. See §10 for a discussion.
Benchmark cost analysis. This is the main part of the cost analysis. This part does not require that the target application is implemented using an MPST-based framework. Assume that we start with a target application, already implemented. The steps of our methodology are the following:

1. Write its global type. Since most of our benchmarks are derived from implementations in other MPST-based tools, this step is straightforward. For non MPST-based implementations, the developer needs to analyse the communication protocol and write it as a global type.

2. Extract the sequential parts. The sequential parts must be extracted as self-contained implementations, that can be run independently of the whole distributed system.

3. Run the profiler on the sequential parts. Our profiler requires multiple input sizes, measures the execution costs of the sequential parts on these input sizes, and performs cubic spline interpolation on the gathered data. Note that the sequential cost is only valid for inputs of sizes that are within the measured range. This part can be omitted when using a static cost analysis, or the cost equations are known and provided manually.

4. Annotate the global type and extract cost equations.

5. Instantiate the cost by feeding the profiling information for both the target architecture and the sequential parts.

8.2 Benchmark Structure

We list and provide a brief explanation of all the benchmarks that we used for the cost models. We used two different hardware configurations for the evaluation. We name them Arch1 and Arch2: Arch1 is a 4-core Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz with hyperthreading, and Arch2 comprises 2 NUMA nodes, 12 cores per node and 62GB of memory, using Intel Xeon CPU E5-2650 v4 @ 2.20GHz chips. Arch2 is an HPC cluster that uses PBS queuing mechanism. We made sure that we consistently selected the same hardware for every execution. In the remainder of this section, we will specify whether the benchmarks were run on Arch1 or Arch2.

We used the benchmarks as defined in the different sources from where we took the source code. Overall, we used averages of > 50 repetitions for benchmarks with large computation costs, and linear regression (95% CI) for smaller (micro-benchmarks such as ping-pong, all-to-all, etc).

Microbenchmarks. Recursive Ping-Pong is the recursive ping-pong example. We run both the Scala benchmark ($pp$-akka) from the Savina benchmarking suite [Imam and Sarkar 2014], and the OCaml version taken from Imai et al. [2020] on Arch1, on three different transports ($pp$-ev, $pp$-lwt, $pp$-ipc-$n$). Since the cost of sending in the $ipc$ transport depends on the input size, we use $n$ to differentiate different runs of this benchmark with different input sizes. We introduced an arbitrary computation to the Scala version to increase the local computation costs. Thread Ring ($ring$) is the Scala version from Imam and Sarkar [2014], both with and without asynchronous message optimisations, on Arch1. Counting Actor ($count$) is a benchmark with two actors, one of which counts the number of messages received from the other. This is the Savina microbenchmark Imai et al. [2020] on Arch1. One-to-All, All-to-One and All-to-All: we use the Go one-to-all, all-to-one and all-to-all Go implementations ($1a$, $a1$ and $aa$) in [Castro et al. 2019], all run on Arch1.

Concurrency Benchmarks. Two-Buyer Protocol ($twobuy$). We use an F# implementation taken from Zhou et al. 2020, and extracted into OCaml. Sleeping Barberx, Dining Philosophers and Cigarette Smoker ($barb$, $dphil$, $csmok$). These are the Savina Akka benchmarks in [Imam and Sarkar 2014], run on Arch1. K-Nucleotide, Spectral-Norm and Regex-DNA ($kn$, $sn$, $dna$). These benchmarks are Go implementations taken from [Castro et al. 2019], based on the Computer Language Benchmarks Game, and use different combinations of scatter, gather, choices and recursion.
Parallel Algorithms. All these benchmarks were run on Arch2, and they were taken from two sources: Ng et al. [2015] (NBody, linear equation solver, wordcount and adpredictor) and Castro-Perez and Yoshida [2020b] (dot product, fast fourier transform and mergesort).

Ng et al. [2015]. The work by Ng et al. [2015] has implemented representative parallel benchmarks from [Asanovic et al. 2009]. NBody (nb) is a 2D NBody simulation in C+MPI which is implemented as a thread ring with asynchronous communication optimisations. Linear equation solver (ls) is parallelised using a wraparound mesh. Similarly to the NBody example, we required the extension with asynchronous communication optimisation. WordCount (wc) and AdPredictor (ap) are parallelised using map-reduce.

Double-Buffering Algorithm. (dbuf) [USENIX 2020] is a well-known technique for increasing the throughput of a device that has two buffers. To accurately represent a double-buffering protocol, we use CAMP’s extension with asynchronous message optimisations. We show the protocol below, using participants p for source, q for sink and r for the service:

\[
\text{rp}!\{r_1\}.\text{rp}!\{r_2\}.\mu X.\text{pr} ?\{r_1\}.\text{p} \rightarrow r\{s_1\}.\text{q} \rightarrow r\{t_1\}.\text{r} \rightarrow q\{u_1\}.\text{rp}!\{r_1\}.\text{pr} ?\{r_2\}.\text{p} \rightarrow r\{s_2\}.\text{q} \rightarrow r\{t_2\}.\text{r} \rightarrow q\{u_2\}.\text{rp}!\{r_2\}. X
\]

Fig. 6 illustrates this protocol. Suppose a streaming service with two buffers (A and B), a source (left) and a sink (right). First (a), buffer A is ready to copy an element (message r₁), and so it notifies the source. Then (b), an element is copied into A (message s₁). Meanwhile, both the sink and B can notify the service and the source (respectively) that they are ready to copy (messages r₂ and t₁). This implies that, next (c), both the service and the sink can copy an element in parallel (messages s₁ and u₁). Note that using a single buffer, this would not be possible, since we would risk overwriting the buffer before the sink copied it. In the next iteration (d), we can swap the buffers, and repeat the process. By swapping the buffers, both the service (buffer A) and the sink can notify that they are ready, even if data is still being copied to buffer B (messages r₂ and t₂). Finally (e), buffer A and the sink can copy the respective next elements, again in parallel (messages s₂ and u₂).

Remark 8.1 (Double-buffering). (1) Definition 6.1 does not directly check the asynchronous subtyping of local types from the above global type as our rules do not include unrolling recursive global types (to obtain the decidability result). However we can apply any (sound) asynchronous subtyping relation from the literature since the cost calculation does not related to well-formedness of global types. For example, local types that behave as the projections of this global type are known as deadlock-free [Mostrous et al. 2009; Yoshida et al. 2008]. (2) The syntax of the global type in [Mostrous et al. 2009] uses the explicit channels in global types. We translated them to corresponding labels, which does not affect the cost calculation, and our end-point implementation is essentially as identical as one in [Yoshida et al. 2008].

Castro-Perez and Yoshida [2020b]. Mergesort (ms) follows a divide-and-conquer protocol. Fast Fourier Transform (fft) is the Cooley-Tukey fast-fourier transform algorithm, implemented
values can be computed in constant time from the results of the same two recursive calls. The complete algorithm for a size-8 is illustrated by the diagram from Fig. 7(b). The global type in Fig. 7(c) shows the resulting global type, in terms of the indices of the participants that need to communicate. We use keyword foreach to represent that the body must be expanded for all natural numbers that satisfy the condition (similarly to parameterised MPST [Deniélou et al. 2012]). CAMP uses a recursive definition that expands into a butterfly of the required size. Fig. 7(d) shows the (abstract) code of our implementation for participants 0 and 1. We show the high-level structure, in terms of send and receive. Suppose that participants receive as initial value $\sigma$, the deinterleaving of the input vector. Then, they all start applying a sequential fft, and communicate the result to the appropriate participants. Then, they apply the necessary addition and subtraction to compute their part of the result, and communicate it accordingly.

8.3 Discussion of Predicted Execution Times

Fig. 8 shows a comparison, for each benchmark, of the execution times compared with the predictions by our cost models. For most of our examples, we get predictions with $< 15\%$ of error.
Examples include `pp-ipc-4`, `a1-2`, `ring`, `count`, `nb-32`, `wc-2`, `ap-2`, `ap-64`, and `dbuff`. We observe that the worst predictions are those of the microbenchmarks, with very small execution times. Here, communication costs dominate, and are repeated a large number of times. With such small costs, a small error is amplified after a large enough number of iterations. An example of this is `ring`, that is a recursive ring protocol that is run for $10^5$ iterations.

When we consider examples with larger local computation costs, most of the predictions are with less than 10% error. There are a small number of examples above than 10% where errors in the prediction are due to factors that CAMP’s cost models do not take into account, such as scheduler costs, cost of thread creation, or resource contention such as shared caches. These details that the cost models do not take into account can also explain why, in some cases, the cost models do not predict an upper bound of the cost, since the real executions include slowdowns due to these factors. Note, however, that CAMP offers a quick and static first assessment of the performance behaviour of concurrent and distributed systems which use different transports and topologies, without the need to deploy or profile the application.

Fig. 8. Predicted vs real execution times: PP = Ping-Pong, AA = All-to-All, S = Scatter, G = Gather, SG = Scatter-Gather, CR = choice with recursion, MR = MapReduce, D&C = parallel divide and conquer.
Asynchronous Communication Optimisations. Algorithms \texttt{fft}, \texttt{dbuff}, \texttt{nb} and \texttt{ring} all rely on asynchronous communication optimisations. Both \texttt{fft} and \texttt{dbuff} require to be specified using this extension. For \texttt{ring}, we take measurements to compare the optimised and unoptimised global types. We can observe a speedup in the execution of the protocol that is predicted by the cost models, which is consistent with Theorem 6.3.

9 RELATED WORK

Resource Analysis and Session Types. Das et al. [2018b] combine session types with amortised resource analysis in a linear type system, to reason about resource usage of message-passing processes, but their work focuses on binary sessions in a linear type system, while we focus on multiparty session types, and the global execution times of the protocol. Das et al. [2018a] extend a system of binary session types in a Curry-Howard correspondence with intuitionistic linear logic [Caires and Pfenning 2010; Caires et al. 2016] with temporal modalities \texttt{next}, \texttt{always}, and \texttt{eventually}, to prescribe the timing of the communication. A fundamental difference with our work is that Das et al. [2018a] require the introduction of delays into the processes, to match the specified cost. In our case, the processes are left unmodified, and the cost is computed from the protocol descriptions. Finally, their work are limited to theory, while our work are readily applied to real use cases.

Asynchronous Communication Optimisation. The first idea of asynchronous communication optimisation was found in Scribble [scribble authors 2008] where a multiparty financial protocol with message ordering permutations is informally described. Later this idea was formalised as \textit{asynchronous session subtyping} for the \(\pi\)-calculus [Chen et al. 2017, 2014; Mostroux and Yoshida 2009, 2015; Mostroux et al. 2009] and its denotational properties were studied in [Demangeon and Yoshida 2015; Dezani-Ciancaglini et al. 2016]. Concurrently, because of the need of asynchronous optimisation in multiparty protocols, several applications inspired by asynchronous subtyping have been developed in Java [Hu 2017], C [Yoshida et al. 2008] and MPI-C [Ng et al. 2015, 2012], but without any formal theories. Recently, this subtyping relation was found undecidable for \textit{binary} session types [Bravetti et al. 2017, 2018; Lange and Yoshida 2017] and its sound algorithm for binary session communicating automata was proposed in [Bravetti et al. 2019]. We have implemented a different and more practical decidable optimisation relation based on asynchronous subtyping for multiparty session types, recently proposed in [Ghilezan et al. 2021]. None of the above work has (1) developed a formal cost theory which can justify the optimisation; and (2) measured and compared the cost of optimised/unoptimised applications with a formal justification. CAMP is the first framework which (1) proposes a formal cost theory with asynchronous optimisation (Theorems 6.2 and 6.3) and (2) justifies the optimisation cost against real benchmarks using (1).

Timed Session Types. The notion of \textit{time} has been introduced to session types [Bartoletti et al. 2017; Bocchi et al. 2015, 2019, 2014], to account for protocols that require time specifications, originated from communicating timed automata (CTA) [Krčál and yi 2006]. Session types and the \(\pi\)-calculus processes have been related in terms of static typing [Bocchi et al. 2019, 2014], or timed session types are linked with compliments relations [Bartoletti et al. 2017] or CTA [Bocchi et al. 2015]. Among them, [Bartoletti et al. 2017; Bocchi et al. 2019] are limited to binary or server-client session types. All of the above works are theoretical only, while the work in [Bocchi et al. 2014] was applied to the runtime monitoring in Python [Neykova et al. 2017]. The main difference is that the above timed session types focus on ensuring that deadlines or time constraints are satisfied. In contrast, our work does not enforce any time constraints, since we are interested on the static estimation of execution costs, but not on enforcing that timeouts and deadlines are respected.

Type-Based and Amortised Cost Analysis. Handley et al. [2019] use refinement types to reason about efficiency, cost, of Haskell programs, but they do not consider concurrency or parallelism. Sized types [Hughes et al. 1996] are one of the successful techniques for cost analysis of programs.
[Avanzini and Dal Lago 2017; Portillo et al. 2002; Vasconcelos 2008]. Most of the uses of sized types do not deal with concurrency and distribution. Exceptions are [Gimenez and Moser 2016], that address space and space-time complexity of parallel reductions of interaction-net programs using sized and scheduled types, but they do not address message-passing and distributed environments. The work [Hoffmann and Shao 2015] extends earlier amortised cost analyses [Hoffmann et al. 2012] to parallel reductions. Their work focuses on parallel functional programs with explicit parallel composition, but does not address message-passing. To our best knowledge, none of the work above addresses the cost of message-passing constructs or distributed environments.

10 CONCLUSIONS AND FUTURE WORK
We have presented CAMP, a framework for statically predicting the cost, execution times, of concurrent and distributed systems. CAMP augments global types from the theory of multiparty session types with local computation costs, and its trace semantics is extended with local computation observable actions. We have developed a way to extract cost equations from these instrumented protocol descriptions, that we can use for estimating upper-bounds of the execution times required by the participants of a protocol. CAMP can be used to predict the latency, i.e. the execution times that the participants of a protocol will require, on average, per iteration of the protocol. Furthermore, we extended CAMP to address asynchronous communication optimisation. CAMP’s cost analysis on top of multiparty session types gives us several benefits. Firstly, we can use global types to reason about both correctness and performance of concurrent and distributed systems. Secondly, the cost analysis can be readily applied and integrated into any MPST framework. Thirdly, it can be used in non-session-based concurrency benchmarks by simply providing MPST protocols. It suffices to describe the global type instrumented with cost, and instantiate the derived cost equations with measured or estimated communication latencies, and local computation costs. And, fourthly our prototype accounts for CPU/CORE availability of the target hardware.

CAMP addresses two main concerns when estimating execution costs of concurrent and distributed systems: communication overheads, and synchronisation. Although these factors are a main source of inefficiency, there are more that we still do not take into account, such as the cost of starting new threads, the cost of context switching/scheduling, or the cost of resource contention such as shared caches [Lea 1997]. We plan to study how to extend CAMP to take such factors into account as future work. CAMP considers distributed systems comprised of multiple nodes, each of which with a number of CPUs/cores. We plan to extend CAMP’s hardware descriptions to consider heterogeneity, e.g. CPU clusters, FPGAs, etc. CAMP’s cost models take the maximum cost of the different possible branches in a protocol. This is sufficient to compute a worst-case execution time of non-recursive protocols. However, we can extend our costs to take into account the weight of different branches, so that our cost models would compute an average cost based on the probability to take the different branches. Moreover, since communication latencies may not be very predictable, we plan to study the extension of CAMP to use probabilistic cost estimations. Finally we plan to study the development of a performance analysis tool for existing code, based on the inference or extraction of the communication protocol followed by non-session-typed implementations such as [Gabet and Yoshida 2020; Lange et al. 2018; Ng and Yoshida 2016].

ACKNOWLEDGMENTS
We thank the OOPSLA reviewers for their careful reviews and suggestions; and Lorenzo Gheri and Fangyi Zhou for their comments. Francisco Ferreira and Keigo Imai helped testing our artifact submission. The work is supported by EPSRC EP/T006544/1, EP/K011715/1, EP/K034413/1, EP/L00058X/1, EP/N027833/1, EP/N028201/1, EP/T006544/1, EP/T014709/1 and EP/V000462/1, and NCSS/EPSRC VeTSS.
REFERENCES


Gadi Taubenfeld. 2006. Synchronization algorithms and concurrent programming. Pearson Education.


